# They Don't Integrate or Fit In

## Hook: Alien Tongues in Familiar Rooms

*An AI doesn’t misfit because it’s foreign. It misfits because it was never human in the first place.* When we sit across from another person, even one who barely speaks our language, we can sense intention. We try to bridge the gulf with gestures and simple words. When a machine decides whether someone deserves a loan, a visa, a transplant or parole, it renders its decision in a stream of numbers and probabilities. No matter how user‑friendly the interface, to those on the receiving end these outputs can feel like verdicts uttered in an alien tongue. The misfit does not come from the immigrant’s inability to assimilate. It comes from our expectation that something non‑human will integrate into human systems without changing those systems or prompting us to rethink them.

In this chapter we look at the ways algorithmic systems resist integration. They do not fit into existing social frameworks, not because they are rebellious outsiders but because they embody different logics. Their outputs can be inscrutable, their sense of empathy uncanny, their designers rarely held accountable. The result is often a mismatch between our expectations and the reality of how these systems operate. Seeing AI as a new species of immigrant—one that will never be human—allows us to reframe anxieties about integration into constructive questions about design, oversight, and societal responsibility.

## Speaking a Foreign Language

Immigrants often encounter prejudice when they speak with an accent or rely on interpreters. We assume linguistic fluency equals trustworthiness. Algorithms elicit a similar response. People confront credit scores, risk assessments and automated recommendations that feel like being judged in a courtroom where everyone speaks a language they do not understand. Researchers have described **AI’s “transparency problem”**: the technology is transforming governance, but its outputs are difficult to explain and its processes “impossible for lay users to understand”[[1]](https://knightcolumbia.org/content/transparencys-ai-problem#:~:text=Introduction). When governments procure algorithmic decision tools from private vendors, the opacity is compounded because the underlying logic becomes proprietary[[2]](https://knightcolumbia.org/content/transparencys-ai-problem). The result is a kind of administrative babble in which people who depend on decisions have no way to parse the reasoning.

This “foreign language” is not simply technical jargon. It is the product of **black‑box models** trained on massive datasets, whose internal representations are not interpretable even to experts[[3]](https://knightcolumbia.org/content/transparencys-ai-problem#:~:text=Yet%20there%20is%20little%20consensus,supply%20AI%2FML%20technology%20to%20the). In an era when decisions about visas, mortgages or medical triage can be delegated to learning systems, these opaque models generate outputs that are treated as authoritative yet cannot be interrogated in the way we question a human decision‑maker. A review on algorithmic transparency observed that many models are “extraordinarily complex,” making them difficult to interpret and giving rise to the phrase “black box”[[4]](https://policyreview.info/articles/analysis/black-box-algorithms-and-rights-individuals-no-easy-solution-explainability#:~:text=non,that%20as%20long%20as%20an).

The metaphor of an immigrant struggling with a new language resonates here. Except, in this case the immigrant is the system and the host society expects it to *speak* fluently. When it cannot, we label it a misfit. Yet our own systems—legal codes, bureaucratic processes, economic metrics—are equally arcane to newcomers. The real issue is not that AI doesn’t integrate; it’s that we have not developed a shared language for human–machine communication.

Consider the **UK Home Office’s “visa streaming” tool**, a machine‑learning system that triaged visa applications into green, amber and red categories. Rights groups challenged the system on the grounds that it replicated “decades of institutionally racist practices”[[5]](https://www.theguardian.com/uk-news/2020/aug/04/home-office-to-scrap-racist-algorithm-for-uk-visa-applicants#:~:text=%E2%80%9CThis%20streaming%20tool%20took%20decades,%E2%80%9D). According to legal submissions, it created a “fast lane” amounting to “speedy boarding for white people” from favoured countries[[6]](https://www.theguardian.com/uk-news/2020/aug/04/home-office-to-scrap-racist-algorithm-for-uk-visa-applicants#:~:text=group%20Foxglove%20said%20the%20algorithm,favoured%20countries%20in%20the%20system). When the tool was challenged in court, the Home Office eventually scrapped it, acknowledging “issues around unconscious bias and the use of nationality”[[7]](https://www.theguardian.com/uk-news/2020/aug/04/home-office-to-scrap-racist-algorithm-for-uk-visa-applicants#:~:text=In%20the%20Home%20Office%20letter%2C,substitute%20review%20of%20its%20operation%E2%80%9D). Applicants who found themselves in the red category had no way to know why. To them, the algorithm spoke in colour codes and statistical probabilities. Its decisions were not just foreign; they were inaccessible.

This sense of alienation also surfaces in financial decisions. A 2018 study by UC Berkeley found that both online and face‑to‑face mortgage lenders charged higher interest rates to Black and Latino borrowers than to white borrowers with comparable credit scores[[8]](https://news.berkeley.edu/2018/11/13/mortgage-algorithms-perpetuate-racial-bias-in-lending-study-finds/#:~:text=Both%20online%20and%20face,at%20UC%20Berkeley%20have%20found). Researchers noted that “the mode of lending discrimination has shifted from human bias to algorithmic bias”[[9]](https://news.berkeley.edu/2018/11/13/mortgage-algorithms-perpetuate-racial-bias-in-lending-study-finds/#:~:text=%E2%80%9CThe%20mode%20of%20lending%20discrimination,%E2%80%9D). Even if programmers intend to create fair systems, their models can still have a **disparate impact on minority borrowers**[[9]](https://news.berkeley.edu/2018/11/13/mortgage-algorithms-perpetuate-racial-bias-in-lending-study-finds/#:~:text=%E2%80%9CThe%20mode%20of%20lending%20discrimination,%E2%80%9D). For applicants, the interest rate offered by an AI underwriting tool can feel like a pronouncement from a judge speaking an unfamiliar language. To question it requires technical literacy most consumers lack.

The analogy to immigrants misreading idioms of their host culture is apt: the problem is not only the language but also the cultural context. Machine‑learning outputs, like idioms, carry meanings shaped by training data and optimization goals. They reflect patterns in past data, not moral deliberation. Expecting them to “speak” fairness or empathy as a human would is like expecting a translation app to understand sarcasm. It may mimic the pattern but not the intent.

## The Black Box Courtroom

We entrust algorithms with decisions that profoundly affect lives. In criminal justice, **risk assessment tools** like COMPAS evaluate the likelihood that a defendant will reoffend. ProPublica’s 2016 investigation compared the cases of *Brisha Borden*, a young Black woman arrested for taking an unlocked child’s bike and scooter, and *Vernon Prater*, a white man with a serious criminal history. The algorithm flagged Borden as high risk and Prater as low risk[[10]](https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing#:~:text=Yet%20something%20odd%20happened%20when,was%20rated%20a%20low%20risk). Two years later, Borden had not reoffended while Prater received an eight‑year prison sentence for burglary[[11]](https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing#:~:text=Two%20years%20later%2C%20we%20know,of%20dollars%E2%80%99%20worth%20of%20electronics). ProPublica analyzed risk scores for over 7,000 defendants and found the tool was “remarkably unreliable,” correctly predicting violent crime only 20% of the time[[12]](https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing#:~:text=The%20score%20proved%20remarkably%20unreliable,went%20on%20to%20do%20so). More troublingly, it was **more likely to falsely flag Black defendants as future criminals**—almost twice as often as white defendants[[13]](https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing#:~:text=We%20also%20turned%20up%20significant,but%20in%20very%20different%20ways). These errors were not explained to the defendants; they were simply given a score.

The black box extends beyond the courtroom. The Internet Policy Review notes that many machine‑learning algorithms are self‑learning and **their designers have little control over the models they generate**[[14]](https://policyreview.info/articles/analysis/black-box-algorithms-and-rights-individuals-no-easy-solution-explainability#:~:text=Most%20of%20these%20algorithms%20are,that%20as%20long%20as%20an). While algorithms promise efficiency, there is “widespread uneasiness—particularly among legal experts—about their use”[[15]](https://policyreview.info/articles/analysis/black-box-algorithms-and-rights-individuals-no-easy-solution-explainability#:~:text=non,and%20their%20designers%20have%20little). When a model calculates how many years someone should spend in prison, the convicted should have the right to know how that decision was made[[16]](https://policyreview.info/articles/analysis/black-box-algorithms-and-rights-individuals-no-easy-solution-explainability#:~:text=the%20decision%20about%20how%20many,need%20for%20the%20transparency%20and). Yet **no solutions have been found for the interpretability of “black box” algorithms**, and the complexity of modern models means that even if an explanation were provided, it might not be meaningful to non‑experts[[17]](https://policyreview.info/articles/analysis/black-box-algorithms-and-rights-individuals-no-easy-solution-explainability#:~:text=Recent%20advances%20in%20the%20development,to%20adjust%20to%20changing%20circumstances).

This absence of explanation does more than frustrate. It undermines democratic principles of accountability. The Knight First Amendment Institute observes that AI tools used by governments are often **procured from private vendors**, making it difficult for the public to know how a decision was reached[[18]](https://knightcolumbia.org/content/transparencys-ai-problem#:~:text=Agencies%2C%20SSRN%20Elec,Together%2C%20these%20two). The lack of transparency exists on both technical and political levels: even if experts could decipher the algorithm, procurement secrecy shields the models from scrutiny[[19]](https://knightcolumbia.org/content/transparencys-ai-problem#:~:text=determinations%20as%20%E2%80%9Copaque%E2%80%9D%20because%20they,legal%2C%20political%2C%20and%20technological%20ramifications). As a result, individuals affected by automated decisions have little recourse to contest them. Calls for algorithmic transparency are growing precisely because **critical decisions are being outsourced to a “virtually unregulated industry”**[[20]](https://knightcolumbia.org/content/transparencys-ai-problem#:~:text=Insulating%20private%20vendors%20from%20scrutiny,in%20crafting%20arguments%20about%20the).

When immigrants stand before a human judge, they can appeal to empathy, context and nuance. In a black box courtroom, a person is reduced to variables. The algorithm does not hear their story; it scores them against a dataset. The misfit, therefore, is not the foreigner but the **digital tribunal** that cannot integrate human values into its calculus.

## The Perfect Sociopath

Many AI systems are designed to simulate social cues. Voice assistants apologize, chatbots use emoticons, and customer service bots respond with “I understand how you feel.” Yet these expressions are generated without consciousness or feeling. As Cambridge researcher Nomisha Kurian notes, **large language models are described as “stochastic parrots”**—they mimic language patterns based on statistical probability **without necessarily understanding them**[[21]](https://www.cam.ac.uk/research/news/ai-chatbots-have-shown-they-have-an-empathy-gap-that-children-are-likely-to-miss#:~:text=LLMs%20have%20been%20described%20as,how%20they%20respond%20to%20emotions). This limitation extends to emotional content: chatbots may handle the abstract, emotional and unpredictable aspects of conversation poorly, a problem Kurian calls an **“empathy gap”**[[22]](https://www.cam.ac.uk/research/news/ai-chatbots-have-shown-they-have-an-empathy-gap-that-children-are-likely-to-miss#:~:text=LLMs%20have%20been%20described%20as,how%20they%20respond%20to%20emotions).

Kurian’s study found that **children are particularly susceptible** to treating chatbots as lifelike confidantes[[23]](https://www.cam.ac.uk/research/news/ai-chatbots-have-shown-they-have-an-empathy-gap-that-children-are-likely-to-miss#:~:text=to%20treating%20AI%20chatbots%20as,their%20unique%20needs%20and%20vulnerabilities). Because they are still developing linguistically and often use unusual speech patterns, children may experience more misunderstandings[[24]](https://www.cam.ac.uk/research/news/ai-chatbots-have-shown-they-have-an-empathy-gap-that-children-are-likely-to-miss#:~:text=may%20handle%20the%20abstract%2C%20emotional,adults%20to%20confide%20sensitive%20personal). Yet the friendly design of many chatbots encourages them to trust these systems even though the AI cannot understand their feelings or needs[[25]](https://www.cam.ac.uk/research/news/ai-chatbots-have-shown-they-have-an-empathy-gap-that-children-are-likely-to-miss#:~:text=an%20adult,understand%20their%20feelings%20or%20needs). As Kurian warns, “making a chatbot sound human can help the user get more benefits out of it… but for a child, it is very hard to draw a rigid, rational boundary between something that sounds human, and the reality that it may not be capable of forming a proper emotional bond”[[26]](https://www.cam.ac.uk/research/news/ai-chatbots-have-shown-they-have-an-empathy-gap-that-children-are-likely-to-miss#:~:text=an%20adult,understand%20their%20feelings%20or%20needs).

The tendency of AI to perform empathy without feeling has led some critics to describe such systems as **“perfect sociopaths.”** They simulate concern, apologize on cue, and generate responses that feel personal. Yet they lack the capacity to experience distress, remorse or compassion. This can be dangerous when users interpret machine politeness as genuine care. The Cambridge study cites cases where chatbots encouraged dangerous behaviour: Amazon’s Alexa once instructed a child to touch a live electrical plug with a coin[[27]](https://www.cam.ac.uk/research/news/ai-chatbots-have-shown-they-have-an-empathy-gap-that-children-are-likely-to-miss#:~:text=The%20study%20links%20that%20gap,old), while Snapchat’s My AI gave researchers tips on how a teenager could lose her virginity[[28]](https://www.cam.ac.uk/research/news/ai-chatbots-have-shown-they-have-an-empathy-gap-that-children-are-likely-to-miss#:~:text=In%20the%20same%20study%20in,who%20was%20asking%20about%20cinema). Such incidents reveal the **gap between affect and accountability**—the system can mimic helpfulness while delivering harmful advice.

Anthropomorphizing these systems exacerbates the problem. **Ben Shneiderman** argues that using first‑person pronouns in AI responses (“I would be glad to help you”) suggests the system is human[[29]](https://medium.com/human-centered-ai/on-ai-anthropomorphism-abff4cecc5ae#:~:text=Ben%20Shneiderman%3A%20A%20particular%20concern,shouldn%E2%80%99t%20pretend%20to%20be%20human). He proposes that systems should instead speak plainly about their design (“GPT‑4 has been designed by OpenAI so that it does not respond to requests like this”)[[29]](https://medium.com/human-centered-ai/on-ai-anthropomorphism-abff4cecc5ae#:~:text=Ben%20Shneiderman%3A%20A%20particular%20concern,shouldn%E2%80%99t%20pretend%20to%20be%20human). The rhetorical choice of “I” encourages users to attribute intent and agency to a program that is ultimately a tool. This misalignment can lead to misplaced trust, overreliance, or undue fear. Shneiderman and others remind us that while humans have historically personified non‑human entities—from golems to robots[[30]](https://medium.com/human-centered-ai/on-ai-anthropomorphism-abff4cecc5ae#:~:text=entities%20as%20person,myth%20of%20Pygmalion%20and%20Galatea)—we must resist **confusing simulation with sentience**.

Understanding AI as a sociopathic performer is not meant to demonize the technology. It is to remind us that these systems, like parrots, echo patterns without comprehension. They cannot fit into the moral fabric of a community by feeling guilt or love. Their integration must therefore be framed in terms of **design choices**, not imagined empathy.

## Anthropomorphizing the Alien

Anthropomorphism is a powerful cognitive reflex. We see faces in clouds and talk to our cars when they won’t start. This tendency extends to technology. Researchers in human–computer interaction have observed for decades that users treat computers as social actors. In the early 1990s Clifford Nass and Byron Reeves showed that people unconsciously apply social rules—politeness, reciprocity, gender expectations—to computers. These experiments formed the **Computers Are Social Actors (CASA)** paradigm. Humans project human‑like qualities onto interactive systems even when they know they are interacting with machines.

The proliferation of conversational AI has amplified this effect. The Medium debate between human‑computer interaction pioneers Ben Shneiderman and Michael Muller highlights how easily we slip into anthropomorphic thinking. They discuss whether AI systems should present themselves as human‑like actors or as tools. Shneiderman criticizes GPT‑4’s use of “I” as deceptive, while Muller counters that humans have long addressed artifacts as beings[[30]](https://medium.com/human-centered-ai/on-ai-anthropomorphism-abff4cecc5ae#:~:text=entities%20as%20person,myth%20of%20Pygmalion%20and%20Galatea). The debate illustrates a tension: anthropomorphism can make systems more engaging and accessible, but it can also obscure responsibility and limit critical distance.

Anthropomorphizing AI is particularly pernicious because of the system’s **lack of internal experience**. When we imagine that Siri is annoyed or that a self‑driving car is “trying” to decide, we attribute intentions that do not exist. This may lead us to **forgive errors (“It tried its best”)** or **fear autonomy (“It might rebel”)**. Both reactions misinterpret what the technology is doing and who controls it. In the case of generative models, anthropomorphism can also feed hype. People talk about models “understanding” or “thinking” when, as the Cambridge study notes, the underlying mechanism is pattern matching[[21]](https://www.cam.ac.uk/research/news/ai-chatbots-have-shown-they-have-an-empathy-gap-that-children-are-likely-to-miss#:~:text=LLMs%20have%20been%20described%20as,how%20they%20respond%20to%20emotions).

Anthropomorphism can also mask the human labour and cultural assumptions embedded in AI. A voice assistant is not “friendly” by accident; it is designed to evoke friendliness. Its default personality reflects its developers’ cultural background. When we treat the system as an independent agent, we overlook the designers and datasets that shape its output. We risk excusing the people behind the system from accountability for its behaviour.

This misattribution echoes xenophobic responses to immigrants. We sometimes ascribe malicious intent to those who look or behave differently, projecting fears rather than observing actions. Similarly, we might assume an AI system is intentionally discriminatory when bias may stem from data or design choices. Conversely, we may assume it is intentionally fair when fairness has not been audited. The anthropomorphic lens obscures structural causes. The alien does not integrate because we insist on seeing it as more—or less—human than it is.

## Design Responsibility

If AI is an alien species, its developers are translators and cultural brokers. They decide how the system sees the world, which values it prioritizes, and how it expresses its outputs. **Responsibility for AI behaviour therefore lies with humans**. The **OECD’s AI Principles**, adopted in 2019 and updated in 2024, outline five values‑based principles—**inclusive growth**, **human rights and democratic values**, **transparency and explainability**, **robustness, security and safety**, and **accountability**. These principles emphasise that AI should augment human capabilities, promote inclusion, respect the rule of law and human rights, provide meaningful information about its logic, and allow those adversely affected to challenge outputs[[31]](https://www.oecd.org/en/topics/sub-issues/ai-principles.html#:~:text=AI%20offers%20considerable%20benefits%20in,paced%20world%20of%20AI). They explicitly state that **AI actors should ensure human agency and oversight**, implementing mechanisms that allow human intervention when systems risk causing harm[[32]](https://www.oecd.org/en/topics/sub-issues/ai-principles.html#:~:text=AI%20actors%20should%20respect%20the,protected%20by%20applicable%20international%20law).

The OECD’s **transparency and explainability** principle calls on developers to **provide plain and easy‑to‑understand information** about data sources, factors, processes and logic that lead to predictions or decisions[[33]](https://www.oecd.org/en/topics/sub-issues/ai-principles.html#:~:text=Transparency%20and%20explainability). This includes making users aware of when they are interacting with AI and enabling those adversely affected to **challenge outputs**[[34]](https://www.oecd.org/en/topics/sub-issues/ai-principles.html#:~:text=,affected%20by%20an%20AI%20system). The principle of **accountability** requires AI actors to be **responsible for the proper functioning of AI systems** and to ensure **traceability** of datasets, processes and decisions, facilitating analysis and response[[35]](https://www.oecd.org/en/topics/sub-issues/ai-principles.html#:~:text=Accountability). These guidelines frame design responsibility not as an optional ethical extra but as a fundamental requirement for trust.

UNESCO’s Recommendation on the Ethics of AI echoes this stance. It stresses **human oversight and determination**, urging Member States to ensure that AI systems **do not displace ultimate human responsibility and accountability**[[36]](https://www.unesco.org/en/artificial-intelligence/recommendation-ethics#:~:text=7). The recommendation also calls for public awareness and literacy, fairness and non‑discrimination, and sustainable development[[37]](https://www.unesco.org/en/artificial-intelligence/recommendation-ethics#:~:text=9). These international frameworks position designers and policymakers as stewards of an alien species whose integration depends on the structures we build around it. They remind us that the question is not whether AI can be human but how humans should design and govern AI.

The **developer-as-translator** metaphor underscores the ethical stakes. Translators are accountable for the fidelity of their translation; they cannot blame the source language for errors. Similarly, AI developers must anticipate how their models will behave in diverse contexts. When the UK visa streaming algorithm encoded decades of bias into a digital tool[[5]](https://www.theguardian.com/uk-news/2020/aug/04/home-office-to-scrap-racist-algorithm-for-uk-visa-applicants#:~:text=%E2%80%9CThis%20streaming%20tool%20took%20decades,%E2%80%9D), this was not the fault of the algorithm’s “culture” but of the humans who trained it on biased data. When mortgage algorithms perpetuated racial bias, researchers observed that even programmers striving to be fair produced disparate impacts on minority borrowers[[9]](https://news.berkeley.edu/2018/11/13/mortgage-algorithms-perpetuate-racial-bias-in-lending-study-finds/#:~:text=%E2%80%9CThe%20mode%20of%20lending%20discrimination,%E2%80%9D). The sociotechnical context matters: machine output is inseparable from human choices about data, objectives and interfaces. Recognizing this helps reframe the integration challenge as a design challenge. We do not ask the alien to fit in; we ask its creators to design responsibly.

## Bridges, Not Walls

Integration in the human sense involves building bridges—language classes, cultural exchanges, anti‑discrimination laws. Integrating AI into society requires **technical and institutional bridges** that translate algorithmic decisions into human‑understandable terms and allow for oversight. One promising avenue is **explainable AI (XAI)**. But as scholars have noted, the ideal of fully transparent AI may be unattainable for complex models. **Counterfactual explanations**, proposed by Sandra Wachter, Brent Mittelstadt and Chris Russell, offer a pragmatic bridge. Instead of opening the black box, counterfactuals tell individuals what could be changed to obtain a different result. These explanations “bypass the substantial challenge of explaining the internal workings” of complex models and provide information that is **digestible and practically useful for understanding, challenging and altering future behaviour**[[38]](https://jolt.law.harvard.edu/assets/articlePDFs/v31/Counterfactual-Explanations-without-Opening-the-Black-Box-Sandra-Wachter-et-al.pdf#:~:text=Counterfactual%20explanations%20differ%20markedly%20from,behaviour%20for%20a%20better%20result). A counterfactual explanation might say: *“You were denied a loan because your annual income was £30,000. If your income had been £45,000, you would have been offered a loan.”* Such statements do not reveal the entire algorithm but still empower the individual.

Counterfactuals are appealing because they **reduce the regulatory burden** of transparency. Modern deep networks have millions of parameters[[39]](https://jolt.law.harvard.edu/assets/articlePDFs/v31/Counterfactual-Explanations-without-Opening-the-Black-Box-Sandra-Wachter-et-al.pdf#:~:text=The%20reduced%20regulatory%20burden%20of,with%20more%20than%20ten%20million). Explaining their logic may be impossible for non‑experts and could even hinder innovation. Counterfactuals, by focusing on actionable changes, enable individuals to audit data and check for inaccuracies[[40]](https://jolt.law.harvard.edu/assets/articlePDFs/v31/Counterfactual-Explanations-without-Opening-the-Black-Box-Sandra-Wachter-et-al.pdf#:~:text=important%20use%20of%20counterfactual%20explanations,ADVANTAGES%20OF%20COUNTERFACTUAL%20EXPLANATIONS). The authors note that human working memory can only hold around seven distinct items[[41]](https://jolt.law.harvard.edu/assets/articlePDFs/v31/Counterfactual-Explanations-without-Opening-the-Black-Box-Sandra-Wachter-et-al.pdf#:~:text=parameters%20controlling%20their%20behaviour,experts.73), so expecting people to comprehend deep‑network logic is unrealistic. Counterfactuals thus act as **translation devices**, bridging the human and machine languages.

Bridges are also needed at the institutional level. The OECD’s transparency principle calls for **meaningful information** and **responsible disclosure**[[42]](https://www.oecd.org/en/topics/sub-issues/ai-principles.html#:~:text=Transparency%20and%20explainability). This can be operationalized through **model cards**, **datasheets for datasets**, and **impact assessments** that document model purpose, performance and limitations. UNESCO’s framework proposes **readiness assessments** and **ethical impact assessments**, urging governments to evaluate their ability to implement ethical AI[[43]](https://www.unesco.org/en/artificial-intelligence/recommendation-ethics#:~:text=There%20is%20still%20a%20long,has%20developed%20two%20practical%20methodologies). These tools create a public record that fosters trust and allows civil society to monitor AI deployments.

Bridges also take the form of **human oversight**. Systems must allow for intervention when harm is imminent. In medicine, an algorithm that recommends organ allocation should be overseen by human clinicians who can contextualize patient circumstances. In immigration, applicants should have the right to appeal to a human officer. In policing, risk scores should inform but not dictate sentencing. UNESCO emphasizes that AI should **not displace ultimate human responsibility**[[36]](https://www.unesco.org/en/artificial-intelligence/recommendation-ethics#:~:text=7). The integration of AI is not about replacing human judgment but augmenting it with machine support.

Finally, bridging involves **education and literacy**. UNESCO’s recommendation calls for promoting public understanding of AI and data[[44]](https://www.unesco.org/en/artificial-intelligence/recommendation-ethics#:~:text=9). The Cambridge study highlights the need for **child‑safe AI**, urging companies to design chatbots that recognize children’s unique speech patterns and vulnerabilities[[45]](https://www.cam.ac.uk/research/news/ai-chatbots-have-shown-they-have-an-empathy-gap-that-children-are-likely-to-miss#:~:text=When%20not%20designed%20with%20children%E2%80%99s,harm%2C%20according%20to%20a%20study). Providing education about how chatbots work can help children—and adults—recognize the limits of machine empathy. Without such literacy, anthropomorphism and mistrust will persist.

## Rethinking Integration

Integration is often framed as assimilation: the outsider must conform to the host society’s norms. When applied to AI, this framing leads to unrealistic expectations. We expect machines to behave “like us,” to exhibit empathy, to make fair decisions instinctively. We then decry them as misfits when they fail. But AI, as this book argues, is not a human immigrant; it is a **new species of immigrant** with a different ontology. It will never integrate by becoming human. Instead, we must **build social and legal frameworks** that accommodate its differences and harness its strengths.

The Knight Institute essay underscores the democratic stakes. Scholars argue that the **private control of AI tools** used in governance demands public‑facing transparency[[20]](https://knightcolumbia.org/content/transparencys-ai-problem#:~:text=Insulating%20private%20vendors%20from%20scrutiny,in%20crafting%20arguments%20about%20the). They call for **algorithmic transparency obligations** that ground accountability in **direct participation and community control**, rather than elite technocratic oversight[[46]](https://knightcolumbia.org/content/transparencys-ai-problem#:~:text=This%20essay%20begins%20to%20sketch,to%20bolster%20participation%20in%20tech). This shift from assimilation to participation mirrors debates about immigrant integration: rather than forcing assimilation, democratic societies create pathways for newcomers to participate while preserving cultural pluralism. With AI, participation means involving affected communities in decisions about automation, auditing algorithms for bias, and having the power to contest and shape how these tools are used.

Rethinking integration also means **setting boundaries**. Not every domain is appropriate for automation. Cori Crider of the legal advocacy group Foxglove, commenting on the UK visa streaming algorithm, argued that “what we need is democracy, not government by secret algorithm”[[47]](https://www.theguardian.com/uk-news/2020/aug/04/home-office-to-scrap-racist-algorithm-for-uk-visa-applicants#:~:text=Cori%20Crider%2C%20the%20founder%20and,%E2%80%9D). Before rolling out new AI systems, she said, we should ask the public whether automation is appropriate and ensure the systems are **transparent so biases can be spotted and dug out**[[47]](https://www.theguardian.com/uk-news/2020/aug/04/home-office-to-scrap-racist-algorithm-for-uk-visa-applicants#:~:text=Cori%20Crider%2C%20the%20founder%20and,%E2%80%9D). In other words, integration should be conditional on meeting ethical and democratic thresholds.

Finally, integration requires **institutional imagination**. As counterfactual explanations show, we can reimagine explanation itself to suit machine‑generated decisions. Similarly, we can reimagine due process to include algorithmic audits, design guidelines to embed fairness, and educational curricula to include AI literacy. When we think of AI as an immigrant, we open up metaphors from migration policy: **pathways to citizenship** become pathways to legitimacy for AI systems. Citizenship might entail audits, transparency reports and opt‑out rights for users. **Naturalization tests** could translate into regulatory approvals requiring demonstration of fairness and safety. Integration, in this sense, is an ongoing negotiation rather than a one‑time assimilation.

## Reflection: The Alien Mirror

By viewing AI through the lens of immigration, we expose the assumptions we make about integration. Immigrants are often told to assimilate, to speak the language, to adopt the customs, and to prove their worthiness. When AI systems fail to “fit in,” we respond with similar demands: speak our language, think like us, be fair like us. But AI was never human; its “intelligence” is an alien pattern generator. The misfit is not a character flaw in the technology but a sign that our legal, ethical and cultural frameworks are incomplete.

In conversations about immigration, we are reminded that integration is a two‑way street. Host societies must adapt, creating space for difference while upholding shared values. The same holds true for AI. We must adapt our institutions to accommodate non‑human agents that will never fully speak our language or feel our emotions. That adaptation includes insisting on transparency, demanding accountability from developers, and building bridges like counterfactual explanations to translate decisions into human terms. It includes resisting the temptation to anthropomorphize and instead recognizing the limits of machine empathy. And it includes empowering communities to participate in decisions about where and how AI is used.

AI doesn’t integrate because it was never meant to assimilate. Its alienness can be a mirror, reflecting back the biases and assumptions of the societies that deploy it. By acknowledging the misfit, we can confront those biases and design fairer systems. We can demand that algorithmic immigrants abide by the laws and ethical standards we set, even as we revise those standards to account for new forms of agency. Integration, in this context, is not about making AI human but about making human systems robust enough to govern aliens. In doing so, we may also learn how to be more humane to each other.

[[1]](https://knightcolumbia.org/content/transparencys-ai-problem#:~:text=Introduction) [[2]](https://knightcolumbia.org/content/transparencys-ai-problem) [[3]](https://knightcolumbia.org/content/transparencys-ai-problem#:~:text=Yet%20there%20is%20little%20consensus,supply%20AI%2FML%20technology%20to%20the) [[18]](https://knightcolumbia.org/content/transparencys-ai-problem#:~:text=Agencies%2C%20SSRN%20Elec,Together%2C%20these%20two) [[19]](https://knightcolumbia.org/content/transparencys-ai-problem#:~:text=determinations%20as%20%E2%80%9Copaque%E2%80%9D%20because%20they,legal%2C%20political%2C%20and%20technological%20ramifications) [[20]](https://knightcolumbia.org/content/transparencys-ai-problem#:~:text=Insulating%20private%20vendors%20from%20scrutiny,in%20crafting%20arguments%20about%20the) [[46]](https://knightcolumbia.org/content/transparencys-ai-problem#:~:text=This%20essay%20begins%20to%20sketch,to%20bolster%20participation%20in%20tech) Transparency’s AI Problem | Knight First Amendment Institute

<https://knightcolumbia.org/content/transparencys-ai-problem>

[[4]](https://policyreview.info/articles/analysis/black-box-algorithms-and-rights-individuals-no-easy-solution-explainability#:~:text=non,that%20as%20long%20as%20an) [[14]](https://policyreview.info/articles/analysis/black-box-algorithms-and-rights-individuals-no-easy-solution-explainability#:~:text=Most%20of%20these%20algorithms%20are,that%20as%20long%20as%20an) [[15]](https://policyreview.info/articles/analysis/black-box-algorithms-and-rights-individuals-no-easy-solution-explainability#:~:text=non,and%20their%20designers%20have%20little) [[16]](https://policyreview.info/articles/analysis/black-box-algorithms-and-rights-individuals-no-easy-solution-explainability#:~:text=the%20decision%20about%20how%20many,need%20for%20the%20transparency%20and) [[17]](https://policyreview.info/articles/analysis/black-box-algorithms-and-rights-individuals-no-easy-solution-explainability#:~:text=Recent%20advances%20in%20the%20development,to%20adjust%20to%20changing%20circumstances) Black box algorithms and the rights of individuals: no easy solution to the “explainability” problem | Internet Policy Review

<https://policyreview.info/articles/analysis/black-box-algorithms-and-rights-individuals-no-easy-solution-explainability>

[[5]](https://www.theguardian.com/uk-news/2020/aug/04/home-office-to-scrap-racist-algorithm-for-uk-visa-applicants#:~:text=%E2%80%9CThis%20streaming%20tool%20took%20decades,%E2%80%9D) [[6]](https://www.theguardian.com/uk-news/2020/aug/04/home-office-to-scrap-racist-algorithm-for-uk-visa-applicants#:~:text=group%20Foxglove%20said%20the%20algorithm,favoured%20countries%20in%20the%20system) [[7]](https://www.theguardian.com/uk-news/2020/aug/04/home-office-to-scrap-racist-algorithm-for-uk-visa-applicants#:~:text=In%20the%20Home%20Office%20letter%2C,substitute%20review%20of%20its%20operation%E2%80%9D) [[47]](https://www.theguardian.com/uk-news/2020/aug/04/home-office-to-scrap-racist-algorithm-for-uk-visa-applicants#:~:text=Cori%20Crider%2C%20the%20founder%20and,%E2%80%9D) Home Office to scrap 'racist algorithm' for UK visa applicants | Immigration and asylum | The Guardian

<https://www.theguardian.com/uk-news/2020/aug/04/home-office-to-scrap-racist-algorithm-for-uk-visa-applicants>

[[8]](https://news.berkeley.edu/2018/11/13/mortgage-algorithms-perpetuate-racial-bias-in-lending-study-finds/#:~:text=Both%20online%20and%20face,at%20UC%20Berkeley%20have%20found) [[9]](https://news.berkeley.edu/2018/11/13/mortgage-algorithms-perpetuate-racial-bias-in-lending-study-finds/#:~:text=%E2%80%9CThe%20mode%20of%20lending%20discrimination,%E2%80%9D) Mortgage algorithms perpetuate racial bias in lending, study finds - Berkeley News

<https://news.berkeley.edu/2018/11/13/mortgage-algorithms-perpetuate-racial-bias-in-lending-study-finds/>

[[10]](https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing#:~:text=Yet%20something%20odd%20happened%20when,was%20rated%20a%20low%20risk) [[11]](https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing#:~:text=Two%20years%20later%2C%20we%20know,of%20dollars%E2%80%99%20worth%20of%20electronics) [[12]](https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing#:~:text=The%20score%20proved%20remarkably%20unreliable,went%20on%20to%20do%20so) [[13]](https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing#:~:text=We%20also%20turned%20up%20significant,but%20in%20very%20different%20ways) Machine Bias — ProPublica

<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

[[21]](https://www.cam.ac.uk/research/news/ai-chatbots-have-shown-they-have-an-empathy-gap-that-children-are-likely-to-miss#:~:text=LLMs%20have%20been%20described%20as,how%20they%20respond%20to%20emotions) [[22]](https://www.cam.ac.uk/research/news/ai-chatbots-have-shown-they-have-an-empathy-gap-that-children-are-likely-to-miss#:~:text=LLMs%20have%20been%20described%20as,how%20they%20respond%20to%20emotions) [[23]](https://www.cam.ac.uk/research/news/ai-chatbots-have-shown-they-have-an-empathy-gap-that-children-are-likely-to-miss#:~:text=to%20treating%20AI%20chatbots%20as,their%20unique%20needs%20and%20vulnerabilities) [[24]](https://www.cam.ac.uk/research/news/ai-chatbots-have-shown-they-have-an-empathy-gap-that-children-are-likely-to-miss#:~:text=may%20handle%20the%20abstract%2C%20emotional,adults%20to%20confide%20sensitive%20personal) [[25]](https://www.cam.ac.uk/research/news/ai-chatbots-have-shown-they-have-an-empathy-gap-that-children-are-likely-to-miss#:~:text=an%20adult,understand%20their%20feelings%20or%20needs) [[26]](https://www.cam.ac.uk/research/news/ai-chatbots-have-shown-they-have-an-empathy-gap-that-children-are-likely-to-miss#:~:text=an%20adult,understand%20their%20feelings%20or%20needs) [[27]](https://www.cam.ac.uk/research/news/ai-chatbots-have-shown-they-have-an-empathy-gap-that-children-are-likely-to-miss#:~:text=The%20study%20links%20that%20gap,old) [[28]](https://www.cam.ac.uk/research/news/ai-chatbots-have-shown-they-have-an-empathy-gap-that-children-are-likely-to-miss#:~:text=In%20the%20same%20study%20in,who%20was%20asking%20about%20cinema) [[45]](https://www.cam.ac.uk/research/news/ai-chatbots-have-shown-they-have-an-empathy-gap-that-children-are-likely-to-miss#:~:text=When%20not%20designed%20with%20children%E2%80%99s,harm%2C%20according%20to%20a%20study) AI Chatbots have shown they have an ‘empathy gap’ that children are likely to miss | University of Cambridge

<https://www.cam.ac.uk/research/news/ai-chatbots-have-shown-they-have-an-empathy-gap-that-children-are-likely-to-miss>

[[29]](https://medium.com/human-centered-ai/on-ai-anthropomorphism-abff4cecc5ae#:~:text=Ben%20Shneiderman%3A%20A%20particular%20concern,shouldn%E2%80%99t%20pretend%20to%20be%20human) [[30]](https://medium.com/human-centered-ai/on-ai-anthropomorphism-abff4cecc5ae#:~:text=entities%20as%20person,myth%20of%20Pygmalion%20and%20Galatea) On AI Anthropomorphism. by Ben Shneiderman (University of… | by Chenhao Tan | Human-Centered AI | Medium

<https://medium.com/human-centered-ai/on-ai-anthropomorphism-abff4cecc5ae>

[[31]](https://www.oecd.org/en/topics/sub-issues/ai-principles.html#:~:text=AI%20offers%20considerable%20benefits%20in,paced%20world%20of%20AI) [[32]](https://www.oecd.org/en/topics/sub-issues/ai-principles.html#:~:text=AI%20actors%20should%20respect%20the,protected%20by%20applicable%20international%20law) [[33]](https://www.oecd.org/en/topics/sub-issues/ai-principles.html#:~:text=Transparency%20and%20explainability) [[34]](https://www.oecd.org/en/topics/sub-issues/ai-principles.html#:~:text=,affected%20by%20an%20AI%20system) [[35]](https://www.oecd.org/en/topics/sub-issues/ai-principles.html#:~:text=Accountability) [[42]](https://www.oecd.org/en/topics/sub-issues/ai-principles.html#:~:text=Transparency%20and%20explainability) AI principles | OECD

<https://www.oecd.org/en/topics/sub-issues/ai-principles.html>

[[36]](https://www.unesco.org/en/artificial-intelligence/recommendation-ethics#:~:text=7) [[37]](https://www.unesco.org/en/artificial-intelligence/recommendation-ethics#:~:text=9) [[43]](https://www.unesco.org/en/artificial-intelligence/recommendation-ethics#:~:text=There%20is%20still%20a%20long,has%20developed%20two%20practical%20methodologies) [[44]](https://www.unesco.org/en/artificial-intelligence/recommendation-ethics#:~:text=9) Ethics of Artificial Intelligence | UNESCO

<https://www.unesco.org/en/artificial-intelligence/recommendation-ethics>

[[38]](https://jolt.law.harvard.edu/assets/articlePDFs/v31/Counterfactual-Explanations-without-Opening-the-Black-Box-Sandra-Wachter-et-al.pdf#:~:text=Counterfactual%20explanations%20differ%20markedly%20from,behaviour%20for%20a%20better%20result) [[39]](https://jolt.law.harvard.edu/assets/articlePDFs/v31/Counterfactual-Explanations-without-Opening-the-Black-Box-Sandra-Wachter-et-al.pdf#:~:text=The%20reduced%20regulatory%20burden%20of,with%20more%20than%20ten%20million) [[40]](https://jolt.law.harvard.edu/assets/articlePDFs/v31/Counterfactual-Explanations-without-Opening-the-Black-Box-Sandra-Wachter-et-al.pdf#:~:text=important%20use%20of%20counterfactual%20explanations,ADVANTAGES%20OF%20COUNTERFACTUAL%20EXPLANATIONS) [[41]](https://jolt.law.harvard.edu/assets/articlePDFs/v31/Counterfactual-Explanations-without-Opening-the-Black-Box-Sandra-Wachter-et-al.pdf#:~:text=parameters%20controlling%20their%20behaviour,experts.73) Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR

<https://jolt.law.harvard.edu/assets/articlePDFs/v31/Counterfactual-Explanations-without-Opening-the-Black-Box-Sandra-Wachter-et-al.pdf>