I Am Not a Robot

The Inverse Turing Test, The Floor of Humanness, and The Flaw of Testing for Humanness

Karen Lancaster[a] and **Kevin Purkhauser**[b]

Abstract. *Seventy-five years after Turing introduced the Imitation Game, public discourse increasingly treats contemporary AI, especially large language models, as having "passed" his test. This paper argues that the more urgent question has quietly inverted: can humans still demonstrate that they are not machines. We call this the Inverse Turing Test (ITT). We trace this inversion through the rise and collapse of CAPTCHA-style systems, which asked humans to solve tasks once difficult for machines but systematically failed because they measured outputs rather than processes and imposed a performance ceiling that machines eventually reached. We argue that a viable ITT must instead be organised around a floor of humanness rather than a ceiling of machine performance. Drawing on four candidate criteria—embodied situatedness, temporal depth, social responsibility, and productive imperfection—we characterise humanness as an ontological condition as well as a broad, inconsistent, context-dependent range of behaviour that eludes purely output-based tests. On this view, what counts as "human" or "machine" is not a fixed boundary but a moving threshold, continually renegotiated as capabilities, contexts, and criteria change.*

1 INTRODUCTION

The Turing Test remains one of the most pivotal and influential tests of machine intelligence to date. No machine has yet definitively passed the Turing Test (TT) as Turing originally described it; all purported "passes" rely on substantially relaxed or altered versions of the Imitation Game's rules, judges, or conversational conditions [1–3]. The TT has nonetheless remained a central reference point in debates about AI, asking whether humans can tell a machine apart from a human when both play the 'Imitation Game' [4] in which a machine attempts to communicate in a way that is indistinguishable from a human. While this was a poignant question in the latter half of the twentieth century, it has lost some of its force in contemporary debates, as large language models approach or exceed human performance on many language-based tasks [3]. Meanwhile, since the turn of this century, the practical focus has quietly inverted: it is no longer only a question of whether machines can pass for humans, but also whether humans can still prove that they are human in an increasingly online world. We call this configuration the Inverse Turing Test (ITT).[1]

In Section 2, we discuss the ITT through three phases of CAPTCHA[2] history, from distorted text recognition in the early 2000s, through object identification in the 2010s, to the collapse of both in the 2020s, as machine learning closed the capability gap. Each phase manifested a version of the ITT, and each failed for the same structural reason: it tested outputs rather than processes, that is, how the output is produced. Once machines[3] could produce the right outputs, the test became futile, since it was no longer possible to tell machines apart from humans. These CAPTCHA iterations were designed around a performance ceiling, testing for something which humans could do but machines could not (for a time, at least). However, humanness is not a high-performance condition, and we suggest that as AI abilities approach and exceed our own, it will become increasingly difficult to distinguish human from machine via a CAPTCHA-like or Turing-like test.

We evaluate four candidate criteria for humanness: embodied situatedness, temporal depth, social responsibility, and productive imperfection. Each is imperfect, and each will become progressively easier for machines to imitate. Yet this erosion is itself revealing: the differences that persist longest between human and machine may be precisely those that no binary test can reliably detect. As the imitable markers converge, what remains of humanness will be the part that resists demonstration.

2 THE INVERSE TURING TEST

The TT has become one of the most famous benchmarks in discussions of AI, capturing the public imagination as a way of determining whether machines can think. The TT—originally called "The Imitation Game" [4]—offered a deceptively simple reframing of a difficult question: rather than asking whether a machine could think, Turing asked whether a machine could imitate a human convincingly enough that a human judge

[a] University of Nottingham, UK;karen.lancaster@nottingham.ac.uk
[b] University of Vienna, Austria;kevin.purkhauser@univie.ac.at

[1] In one sense, this inversion is already latent in Turing's original setup, which defines intelligence through an interaction in which human and machine are mutually liable to be mistaken for one another. What is new here is the practical configuration: humans are now routinely required to prove their non-machineness to algorithmic systems and this shifts the centre of gravity from "can machines pass as human?" to "under what conditions can humans still count as human within such tests?".

[2] CAPTCHA stands for Completely Automated Public Turing test to tell Computers and Humans Apart, and is used in situations where websites need to prevent automated abuse ("spamming") or to verify that a user is human. A CAPTCHA typically appears on online account creation or login pages, as well as online forms or for online purchases.

[3] We use the term "machine" expansively, to include software such as LLMs and AI, as well as more traditional hardware or robotic technologies.

would believe it is in fact another human. In its original form, the TT consists of a human judge engaging in a text-based conversation with two unseen participants—one human and one machine—and attempting to determine which is which [4, Sec. 1]. If the machine constantly misleads various judges into believing it is a human, it is said to have passed the test and demonstrated something at least in the neighbourhood of intelligence.[4]

No machine has ever definitively passed the TT as it was originally formulated. Part of the reason for this is that Turing's description left key details unspecified, making strict adherence to the proposed experiment impossible [5]. The paper did not define the number of attempts a machine had to mislead the judge, the duration of the conversations, nor precise success criteria. As a result, all later trials which claim to have found a technology which "passed" the TT relied on modified or interpreted rules. Any test of machine intelligence which purports to be a TT must therefore involve some decision-making regarding the parameters of the test and how success (or failure) can be determined.

The TT has endured not because it is philosophically watertight, but because it is philosophically productive and provocative. It has generated decades of debate, experimentation, and results which its author likely did not anticipate. It has also generated, somewhat quietly, a set of assumptions that have gone largely unexamined: that human cognition is the unmarked baseline of intelligence, and that the human participant in a TT simply has to be themselves in order to distinguish themselves from the machine. The test also assumes—perhaps optimistically—that humans are discerning and reliable judges of intelligence (or humanlike communication), and it does not account for the possibility that the average human might lack the knowledge or insight to distinguish a machine from a human. In short, the test presupposed confidence in human ability and scepticism about machines' abilities. The asymmetry was clear: machines performed, whereas humans judged.

Over the past 25 years, however, that asymmetry has inverted. Today, in CAPTCHA challenges (discussed further below), behavioural analytics, plagiarism detection systems, and bot-detection algorithms, the salient question is not whether a machine can prove to a human that it passes as human, but whether a human can prove to a machine that *they* pass as human. The machine, using some platform or algorithm, evaluates us to determine whether we appear sufficiently human: suspicion is no longer directed solely at machines; it has extended to humans as well.[5] The rise of AI has brought about a shift in epistemic authority, blurring the lines between human and machine behaviour. As AI systems act autonomously and mimic human reasoning and conversational style, both humans and AI can be treated with scepticism, challenging traditional assumptions about trust and credibility.

In the original TT, the human was structurally exempt from scrutiny. In the ITT—where the human must prove their humanness to a machine—that exemption is removed. Everyone is under suspicion. Humanness is no longer assumed; now it must be demonstrated. And in many of today's verification contexts, the gatekeeper making that determination is not another human, but an automated system, which assumes we are "guilty" of being non-human until proven otherwise.

The normative stakes have also changed. In the original TT, failure meant that the machine had not yet reached a threshold of interest, but neither the machine nor its creators faced sanctions or restrictions. The same is not true of the ITT. If one fails an ITT in the form of a CAPTCHA, one is prevented from creating an account, making an online purchase, or accessing information, as one is assumed to be a bot. If one fails an ITT in the form of a system which (ostensibly) detects AI-generated text, one may face additional scrutiny, reputational damage, or the dismissal of one's work as AI-generated. Unlike in the original TT, in the ITT our humanness is judged by a machine, which becomes the ultimate gatekeeper, determining whether I am human enough to be permitted to do the task I want to do.

There is a further complication worth noting before we proceed. The widespread claim that chatbots have "passed" the TT relies on significantly relaxed conditions: very short conversations, lenient or untrained judges, and none of the sustained probing that Turing's original formulation suggests [3]. Wherever interpretation or decision-making by researchers is possible, one can make choices that increase the likelihood of a putative pass—for example, by choosing a small number of times the machine needs to mislead the human judge, by keeping conversations brief, or by selecting particular kinds of people to be judges or interlocutors. This matters because it means the inversion we are describing is not a response to AI having definitively succeeded at passing the TT; it is a response to something more subtle: a shift in public perception, in practical infrastructure, and in the direction of suspicion—of humans rather than machines. It is therefore an irony that although no machine has been able to pass the TT, it is nevertheless perceived as apt for machines to distinguish between humans and machines. Importantly, the machine does not need to have passed the TT in order for it to become a judge within an ITT. We now ask whether an ITT is fit for purpose; after that, we suggest some of the ways in which humans remain different from machines, while noting that these ways are becoming increasingly difficult to test for.

3 THE CAPTCHA GENEALOGY

CAPTCHA history offers a compressed empirical narrative of the ITT's development. Each phase enacted a version of the ITT. Each iteration was a plausible attempt to solve the growing problem of bot accounts; however, although most CAPTCHA phases worked for a period of time—because they were able to distinguish a bot from a human—each new iteration ultimately faced problems when bots (or their programmers) developed ways to circumvent the tests. Wherever a test exists, people (or AI) can find loopholes, workarounds, and methods to thwart it. Thus, although each generation of CAPTCHA worked at first, over time it became easier to cheat the system, and each one ultimately had to be aban-

[4] Turing did not specify how many judges would need to reach the erroneous conclusion that the machine is human in order for the machine to be said to have passed the test(s). However, given that there is a 50% chance of the judge guessing correctly, one would surely wish to stipulate that a substantial number of judges are misled in order to deem the machine to have passed the TT.

[5] CAPTCHA and similar systems do more than simply catch malicious or automated programs—they also help to instil public confidence that a system is secure and well-managed, reinforcing perceptions of professionalism and trustworthiness.

doned in favour of a new test. The obvious conclusion is not that we need to develop a better CAPTCHA, but that CAPTCHA-type tests themselves are flawed, in that there is a constant arms race as bots and bad actors work toward passing each new generation.

3.1 The 2000s: Doing What Machines Cannot

In the late 1990s and early 2000s, the early internet faced a growing problem: automated programmes—bots—were creating fake accounts, spamming forums, buying event tickets instantaneously, and manipulating online polls [6]. It became necessary to distinguish human users from bots at scale.

CAPTCHA arose in the early 2000s as a response. Its earliest and most widely adopted form showed users distorted text and asked them to identify what was written. The underlying logic was straightforward: humans would be able to identify distorted text, whereas machines would not; thus, if a user correctly identified the distorted text, they passed the CAPTCHA and were permitted to proceed.[6] Although this approach worked for a few years, bots eventually caught up and were able to pass. Something new was needed to distinguish humans from bots.

3.2 The 2010s: Recognising What Machines Cannot

As bots became better at parsing distorted text, the test needed to become more nuanced. The ostensible solution was to shift from textual to visual scene understanding. Systems began showing users photographs divided into grids, asking them to identify which sections contained traffic lights, storefronts, buses, motorbikes, or suchlike. The thought was that scene understanding requires embodied common sense: you need to know what a bus is in a contextual and experiential sense, not merely what it looks like in isolation. Only humans, it was assumed, possessed sufficient ability to parse a photograph and identify its elements in context.

Whilst this assumption was true for a period of time, it did not hold indefinitely. By the mid-2010s, machine learning systems had been trained on exactly these sorts of tasks, partly by using human CAPTCHA responses as training data [7]. The test had, in a precise irony, trained AI to circumvent its own test. Once again, the CAPTCHA was failing, and something new was needed.

3.3 The 2020s: The Checkbox and the Behavioural Layer

By the 2020s, both prior approaches had been largely superseded. The dominant form of human verification became a simple checkbox, often accompanied by the text "I am not a robot." However, unlike its predecessors, in this CAPTCHA the checkbox itself is not the test. The real evaluation happens underneath: behavioural and environmental signals including interaction patterns, timing, and session data are used to determine whether the user is a human or a bot; the visible interface is a decoy [8].

This shift is telling. The move from distorted text to image grids to behavioural analysis reflects a recognition that output-based tests are inherently gameable and structurally vulnerable: once a machine can produce the right answer, the test collapses. The behavioural approach probes not what was produced, but the conditions under which it was produced. This iteration of CAPTCHA is noteworthy because the focus shifted from humans outperforming machines to humans being slower and less accurate than machines. In earlier CAPTCHAs, the ITT was predicated on the idea that humans could do something which machines could not, and humans proved their humanness by adequately performing the task at which a machine would have failed; in this iteration, humans are identified by their slow movement of the cursor (in addition to other background information such as browsing data). What marks us as human is no longer that we excel at something machines cannot do, but rather, that we are imperfect at something which machines excel at.

Yet even in this new form of CAPTCHA, the gap between human and machine is narrowing. Recent work has shown that automated systems are becoming better at passing behavioural CAPTCHAs—not by being more efficient, but by being less so [8]. Humans move the mouse slowly, inaccurately, with hesitation. For a bot to appear human in its mouse movements, it must operate below maximum performance, that is a simulation of limitation. In previous iterations of CAPTCHA, humans demonstrated their humanness by outperforming bots. In its most recent iteration, humans are noticeable because of their imperfections—slowness, inaccuracy, hesitancy—and bots therefore need to underperform in order to pass for human. It is much easier to underperform than to perform beyond one's actual abilities, meaning that a test which notices imperfections as markers of humanness is easily thwarted by bots, which simply need to be slow and inaccurate in order to appear humanlike. Browsing history and other markers of humanness also seem to be things that a bot could simulate.

The structural lesson of the CAPTCHA genealogy is not that we chose the wrong outputs to test for humanness. It is that any output-based test will eventually be closed by capability improvements of machines. For a period, humans appeared human because they were better than machines at particular tasks. At some point in the late 2010s, the logic inverted: humans now appear human not because they outperform machines, but because they are less efficient and accurate than a bot operating at full capacity. But whichever of these features an ITT tests, machines close the gap and become increasingly able to appear human, while humans struggle to distinguish themselves from a bot.

This presents us with the central question: if the ITT cannot be won by humans outperforming machines, what can it be won by?

4 THE ABDUCTIVE QUESTION

One of the most prominent recent candidates for a structural human–machine distinction is abductive reasoning. A recent paper from DeepMind argues that LLMs are struc-

[6] Interestingly, this system also served a secondary purpose: it helped digitise large collections of books for projects such as Google Books, exploiting a task where human perception outperformed machine optical character recognition (OCR) [7].

turally incapable of the kind of inferential leap exemplified by Einstein's development of general relativity—a move from anomalous observations to a radically new conceptual framework not derivable from prior data by standard inferential procedures [9]. In Charles Peirce's tripartite framework of deductive, inductive, and abductive reasoning [10], LLMs are powerful engines of the first two, but unable to perform genuine high-level abduction.

We take this argument seriously as a research programme while noting its complications. Work such as that by Krenn et al on AI-assisted scientific discovery suggests the line between simulated and genuine abduction may be harder to draw than the DeepMind paper implies [11]. More fundamentally, it is not entirely plausible to claim that machines simply cannot—and will *never* be able to—perform abductive reasoning in any form. In fact, LLMs produce inferences to best explanations regularly; the question is whether these inferences involve the kind of creative, embodied, temporally extended leap that Peirce had in mind, or whether they are sophisticated pattern completions that mimic the surface form of abduction without instantiating its process.

We do not resolve this debate, but we do note a further problem that applies even if the abductive gap is real: abduction is not a valid ITT criterion, because not all humans perform well at abductive reasoning tasks. For a test to function as a CAPTCHA or other ITT, it must identify something that all humans—including those with limited reasoning ability—can do.[7] Since a number of humans would fail abductive reasoning tasks, it is not a membership criterion for humanity. Even if it were certain that machines could never pass an abductive reasoning test, one also needs to be certain that humans can pass it—and this is simply not the case.

Given recent advancements in LLMs, and the fact that previous CAPTCHAs have only managed to persist for a few years before AI became able to pass the test, it seems highly plausible that even if it were presently true that machines cannot pass abductive reasoning tests but humans can, such a claim will not continue to be true indefinitely. Therefore, using abduction as yet another iteration of CAPTCHA or Turing-like test is ultimately going to be untenable.

5 THE FLOOR, NOT THE CEILING

The CAPTCHA genealogy reveals a flaw that runs deeper than the choice of task. The early phases of CAPTCHA, and most ITT design more broadly, presupposed a performance ceiling of machines: to prove you were human, you were required to excel at something which a machine struggled to do, such as reading distorted text or recognising scene elements. In each case, humanness was operationalised as a kind of competence, and the test asked whether one could meet a performance standard. The same would be true if an abductive reasoning test were used to distinguish humans from machines. Even the checkbox CAPTCHA of recent years looks for markers of humanness in the way we move the cursor and the speed at which we type, requiring machines to emulate our humanlike outputs and behaviours in order to appear human.

Although this sort of structure, for a limited period, was adequate to distinguish human from machine, it is not future-proof; as we have seen, machines were able to circumvent each CAPTCHA test within a few years of its being constructed. And indeed, if abductive reasoning tests were implemented (even if one could be designed that all humans could do), the chances are high that machines would in time be able to pass such tests. The logic of using a test which examines a user's abilities or apparent behaviours is fundamentally problematic, given that machines are constantly developing.

Consider what humanness actually encompasses. Some humans are brilliant; others struggle with basic literacy. Some days a person is sharp and articulate; other days they cannot remember the word they want or parse simple phrases correctly. Human performance is not a ceiling to be reached. It is a vast, irregular, context-dependent range, bounded at the top by moments of genuine excellence and at the bottom by fatigue, distraction, illness, inattention, ignorance, and the ordinary messiness of embodied life. A present-day test in which one needs to outperform a bot in order to be acknowledged as human is flawed insofar as there is now too much crossover between the range of normal human abilities and the abilities of bots.

From now on a valid ITT should not ask whether you can reach a ceiling. It should ask whether you fall within a range of human capabilities. For this it needs not merely a ceiling, but a floor: a lower bound of recognisably (and ideally uniquely) human traits, rather than an upper bound of performance which requires humans to behave in particular ways to be recognised as humans.

The shift in focus of CAPTCHA tests in the 2010s is curious. To pass the most recent ITTs, a machine cannot simply get better; getting better is exactly what exposes it as a machine. Instead, to seem plausibly humanlike, it must learn to be bounded: inconsistent, slow, occasionally incoherent, variable across time and context, and wrong in ways that are recognisably human rather than randomly distributed. It must simulate limitation [12] as well as intelligence [13], or rather the specific shape of human unintelligence [8]. And limitation, unlike intelligence, has no single ceiling to aim for: it is biographical, contextual, and relational. Such imperfections and limitations are genuinely hard to fake, not only in a single exchange, but across the temporal depth of a real life.

We cannot yet fully specify what that floor looks like. The four criteria we examine in the following section—embodied situatedness, temporal depth, social responsibility, and productive imperfection—each face serious objections when taken individually, and we do not claim otherwise. But the conceptual shift from ceiling to floor is itself the focus of this paper. The difficulty of specifying the floor precisely is not a failure of the argument; it may be the most honest acknowledgement one can make. This paper does not purport to be the final word on what the floor of humanness is, but rather, to open a dialogue about what it could be. What makes something human is a range to be inhabited, and hu-

[7] There are of course some humans who cannot pass a simple CAPTCHA such as identifying distorted text or elements within a photograph, perhaps due to being young children or having cognitive impairments. However, all adults who are able to use the internet to open accounts or make online purchases are able to do these things. The same cannot be said of abductive reasoning tasks.

[8] Wimsatt and Dreyfus are making related but distinct points. Wimsatt is about heuristic distortion as a feature of bounded cognition, Dreyfus is about the irreducibility of embodied know-how.

manness is an ontological status rather than a performance or ability. Moreover, the range itself is not fixed: what distinguishes us from machines does not stay the same from one year to the next. As machine capabilities advance and human behaviour adapts in response, what counts as recognisably human shifts with it. The floor is thus a moving threshold, and any adequate account of human authentication will need to track that movement rather than resolve it once and for all. Rather than being a solved problem, it will be a research programme for future writers.

There is a broader critical point embedded here. The ceiling-based structure of most ITT design is not just a technical error; it reflects a wider cultural assumption that performance is the measure of value, and that the goal of any test is to establish how high something can go. Our floor-based alternative challenges this. It suggests that what makes something human is not its capacity for excellence but its embeddedness in a life that includes failure, inconsistency, and the irreducible texture of ordinary limitation. In this respect, the ITT as we are reframing it is not only a test design problem; it is also a quiet critique of how we have come to think about performance itself.

6 CANDIDATE CRITERIA FOR A FLOOR-BASED ITT

If the goal is to test for membership in the range of humanness rather than proximity to its upper limit, what criteria are available? We propose four candidates, none individually sufficient, but together as a cluster, offering a more defensible basis for human authentication than any single output-based test. Note that these are not being suggested as candidates to distinguish humans from machines in CAPTCHA-type situations—CAPTCHA tests need to be quick and perfunctory, lest they infuriate human users who must pass extensive psychological tests simply in order to purchase an item or buy tickets. Rather, we propose these four criteria not as workable tests but as diagnostic illustrations: each one identifies something genuinely distinctive about human existence, and each one reveals, on inspection, why that distinctiveness resists operationalisation. The pattern across all four is the same: the moment a criterion becomes measurable enough to test for, it becomes measurable enough to simulate. This is not a failure of the criteria. It is the central structural problem of the ITT—and, by extension, of the TT itself. When an ITT defines the abilities which are (supposedly) uniquely human, machines and their programmers can focus on recreating those abilities in order to pass as human in a TT; it is like sharing a practice exam paper and some model responses, insofar as it then enables readers to work towards emulating those responses themselves.

These floor-based features of humanness are therefore not behavioural tests (it would make little sense to test the humanness of a user by seeing just how poorly they can answer a question, since the lower limits of human ability can be easily replicated by machines). Instead, these features are more ontological in nature: they examine not what a human can do, but what a human *is*.

6.1 Embodied Situatedness

Humans are physically located [14]. They are tired, hungry, cold, or distracted. They have sensory memories, proprioceptive habits, and a body that leaves traces in how they think and write. The slight drag in a sentence written late at night, the particular error pattern of someone typing on a phone, the way attention wanders after a long passage—these are not flaws to be corrected, but signatures of a body in the world.

Generally speaking, the sorts of bots which are candidates in a Turing Test or CAPTCHA do not have bodies. They do not have days or nights, fatigue-like states, or the experience of reading something while simultaneously listening to a conversation across the room (indeed, they do not have any experiences at all). They can simulate these things in output, but simulation at the level of a single passage is different from the coherent trace of an embodied life running through a body of text over time.

A floor-based test grounded in embodied situatedness would not ask whether your prose is perfectly formed. It would look for the coherent signature of a body: not random noise, but structured, contextually appropriate imperfection.

However, distinguishing a human from a bot becomes particularly difficult when bodily situatedness is used as a defining criterion within a text output such as a CAPTCHA or TT-like conversation. Features often associated with human embodiment—such as wandering attention or typing mistakes—might appear to signal a human behind the text, yet these markers are not always reliably visible: machines can easily simulate hesitancy, imperfections, spelling or grammatical errors, and many humans are capable of producing polished, error-free prose even under distracting conditions. As a result, such a criterion—although potential evidence of humanness—is not by itself an adequate test.

Moreover, embodied robots muddy the water further. An LLM can produce flawless text instantly, but an embodied robot may be much less able to do so. A robot with dextrous hands which types on a standard keyboard may make occasional typing errors, potentially giving its text a humanlike quality of occasional typographical errors, yet all the while it is AI-powered (and not trying to simulate errors, just as humans do not). A system trained to reproduce the signatures of embodied limitation—hesitation, fatigue, contextual error—does not thereby become embodied, but it becomes harder to distinguish from something that is via a text-based or screen-based interface. So although (at present) it is easy for a human to distinguish a human from a machine via a cursory glance when in person, it is far less straightforward under TT or ITT (such as CAPTCHA) conditions. This means that although bodily situatedness is an important distinguishing feature between humans and machines, testing for it under blind Turing-like conditions would be exceptionally difficult.

6.2 Temporal Depth

Humans are beings in time. They have biographies: a past that actually happened, that shaped them in specific ways, and that left traces they can access imperfectly, with the characteristic distortions of human memory—some things vivid, others vague, others confabulated [15]. They have a future

they are uncertain about, and an ongoing present in which they are situated.

LLMs have a training cutoff and a context window. They do not have a past in the biographical sense. They can produce plausible biographical content, but they cannot be probed with the consistency and contextual specificity that genuine biographical memory produces. A test designed around temporal depth would not ask for facts, but for the texture of remembered experience: not "what did you do last summer?" (to which a bot can easily construct a lie) but the kind of question that only makes sense if you were actually there, in a life that accumulated.

Temporal depth is also what makes inconsistency coherent rather than random. A human who changes their view over months, or who remembers something differently on two occasions, is being humanly inconsistent [16]. A system that produces statistically variable outputs is being randomly inconsistent. Any test which uses human inconsistency and fallibility as a diagnostic feature would need to distinguish between these.

This would be no easy task, however. Although humans possess lived experiences which unfold over time, this is not necessarily something which can be picked out as a feature found in all and only humans. Machines can generate convincing fabricated memories or past experiences if prompted to do so. Moreover, machines can have genuine 'memories' of events or conversations with particular people, or can recall things they have done; whilst it may lack the first-personal experience of a human memory, it is nevertheless a recording of an event which occurred. At the same time, research on human memory shows that people's recollections are often incomplete and inconsistent [15], with substantial variation between individuals—from those who can remember minute details of what occurred on any given day in their lives [17], to those who struggle to recall what they did yesterday [18]. Temporality is part of the human condition, but distinguishing humans from machines purely in reference to it is fraught with difficulty. A system that learns to produce biographically coherent, consistently distorted memory-like outputs which are imperfect in a humanlike way does not thereby have a past, even though it can simulate having one—but it becomes a more convincing imitator of something that does.

6.3 Social Responsibility

Humans are social animals. They exist in social networks that make them responsible. They have names, histories, relationships, and consequences. They can be found, confronted, and held responsible. This is not a cognitive criterion of humanness, but a social and ontological one [19].

No current AI system is responsible in this sense. A system can simulate a voice, a perspective, a relationship, but it cannot be responsible for what it says in the way a person can be, because there is no self that persists across contexts, accrues reputational consequence, and can be called to answer. Social responsibility is, in this sense, a structural difference rather than a capability gap. As such it is a matter of what kind of thing it is.

Although at present, no machines are held responsible for their actions, there are philosophical arguments to suggest that they could or should be at least partly responsible when certain criteria are met [20]. However, the distinction is that humans (over a defined minimum age) have legal and moral responsibility for their actions unless it can be proven otherwise (with reference to diminished capacity or suchlike). With machines and software, the reverse is true: the default assumption is that they do not have any responsibility, and if responsibility is to be assigned to them, it must be robustly philosophically or legally proven. To date, no machine or software has been held morally or legally responsible for its actions.

Thus humans are (at present) distinct from machines not because of their intelligence, but because they exist in a web of consequence: when a human user takes an action, there is someone behind the screen who could be found and held responsible for their actions. The same is generally not true for machines.

The problem, however, is determining whether the entity being conversed with is one which actually has responsibility, or merely claims to. An LLM is quite capable of claiming to have social responsibility and behaving in a way suggestive of a moral perspective; meanwhile, some humans are content to shirk responsibility which is rightfully theirs. Although it may be true that humans are responsible for their actions whereas machines are not, testing for responsibility faces the same problems as the original TT: one cannot test whether an entity *is* responsible; one can only test whether it *claims* to be responsible. A system trained on the markers of responsibility—hedging, consequence-awareness, relational tone—is not necessarily responsible. But the gap between performing responsibility and having it is becoming harder to locate.

6.4 Productive Imperfection

In the 1985 movie *D.A.R.Y.L.*, an android engineered to pass as a normal human boy is perceived by his foster parents to be unsettling not because he fails at anything, but because he is too perfect. Upon hearing this from his friend, the android boy intentionally fails to hit the ball during a baseball match, then swears at his foster parents. He later reflects that he had learned something important about passing for human: sometimes failure is more appropriate than peak performance. This rings true in everyday life: perfection, in human contexts, often provokes suspicion rather than trust. Research in social psychology suggests that minor errors can paradoxically increase a highly competent person's perceived likeability—an effect known as the Pratfall effect [21]. The implication for human authentication is pointed: errors, rather than being mere noise, carry social and relational meaning. A system that performs flawlessly and instantaneously signals its own non-humanness not despite its competence but because of it.

Today's customer service bots are often identifiable because they respond instantaneously with paragraphs of perfectly detailed text: research on chatbot design has found that deliberately delayed responses feel more humanlike precisely because instant responses do not. Human agents hesitate, abbreviate, and make small errors which they correct mid-sentence. Of course, customer service chatbots are not flawless—they can misunderstand questions, struggle with nuanced or incomplete information, and fail to answer novel questions. Nevertheless, the prose which bots do produce is often flawless and instantaneous in a way that human-produced text is not.

The em dash has become a minor emblem of human imperfection in academia and elsewhere. The em dash, which is not easily produced on a standard keyboard, is frequently generated by LLMs as the correct form of punctuation. Humans, by contrast, more often reach for a hyphen or an en dash, which is easier to produce when typing manually. It has become almost a cliché that heavy use of the em dash in written work suggests AI generation [22]. In response, some writers now deliberately replace em dashes with en dashes and introduce minor errors—an awkward phrase here, a misplaced comma there—as signals of human authorship. Errors that were once signs of carelessness have become, perversely, something to cultivate as proxies of humanness.

The logic is not simply that humans make mistakes, but that humans make particular *kinds* of mistakes, in particular distributions, under particular conditions, and that these patterns are recognisable as human in ways that both random noise and perfect output are not. Productive imperfection, properly understood, is not about introducing errors, but about expressing the genuine texture of humanly bounded performance.

The irony is that an LLM can be instructed to underperform as well: to introduce misspellings, use en dashes, and so on. Popular LLMs are currently poor at producing text slowly, and cannot "intentionally" pause before constructing a response; nevertheless, this behaviour could in principle be coded into LLMs if it were shown that users wanted it. But being slow or inaccurate deliberately is a different thing from doing so naturally (as humans do), and the difference may be traceable, especially across time and contexts. A human who is tired writes differently at midnight than at noon. An LLM instructed to simulate tiredness produces a statistical distribution of tiredness-associated features that does not vary with the actual time of day, or with what the system was doing before, because it was not doing anything before. A system instructed to underperform in humanlike ways does not thereby become limited, but limitation, once specified, is exactly what can be engineered into the system such that it can appear to be limited in a humanlike way.

The four features identified above—embodied situatedness, temporal depth, social responsibility, and productive imperfection—are all markers of humanness which distinguish us from bots, but unlike previous iterations of CAPTCHA-style ITTs, they are not grounded on the idea that humans must outperform a machine in order to be recognised as human. Rather, they operate on what we are calling a floor-based system, where humans are recognised as such for basic features of their existence rather than by having abilities which exceed machine abilities. A floor-based ITT thus asks whether a human is in the picture at all, rather than asking whether a human can reach a particular standard or demonstrate a particular ability. It asks whether you carry the coherent signature of an embodied, temporally extended, socially responsible, and imperfect life, with all the messiness that entails. Although it may be difficult for a CAPTCHA or other ITT to test for these things, it is going to become increasingly necessary to do so as the gap between human and machine ability narrows.

A clarification is necessary here. This floor is not uniquely human in the strict philosophical sense: some non-human animals possess embodied situatedness and social responsibility in greater measure than some humans. Nor does it cover every human being in every condition. What it describes is something narrower and more pragmatic: it distinguishes the kind of being that could plausibly be operating an internet connection, making a purchase, and so on. That is a communicative and contextual boundary rather than an ontological one, and it is the appropriate boundary for the ITT as we are framing it.

A second clarification concerns testability. The four criteria we have outlined cannot be evaluated in the ten to twenty seconds that a CAPTCHA-style interaction allows, and we do not claim otherwise. They are criteria for extended, probed conversation in the spirit of Turing's original Imitation Game, not for quick verification at a login screen. The speed constraint of practical authentication is a genuine limitation, and one that remains unsolved.

There is, however, a small irony worth noting. The frustration a human feels at being delayed by a verification system—the impatience, the sense that their time is being wasted—is itself a recognisably human response that no current system experiences. If anything, our anger at a prolonged CAPTCHA test may be among the more reliable signals it could use (though in time, a machine could simulate that, too).

The machine's problem, in this frame, is not that it lacks intelligence; rather, that it lacks limitation in the right sense. What it needs to simulate is not artificial or random imperfections, but the structured, contextually coherent, biographically-grounded imperfection of a particular life. That is a harder target, and perhaps permanently harder, because it is a moving one.

7 THE AGENT-TO-AGENT CONTEXT

We situate these arguments in a broader shift that is currently underway: the move toward agent-to-agent AI communication, in which AI systems increasingly interact with each other rather than with human interlocutors. In such an environment, the question of origin has become structurally important in a new and interesting way. Transparency of origin—the simple declaration "I am a human" or "I am a machine"—is no longer a reliable norm. Verification of origin is replacing it.

This institutionalises the ITT in a way that goes beyond security applications, extending into education, employment, journalism, creative authorship, personal relationships, and political discourse. In all of these domains, there are situations in which human-generated text is felt to be more authentic, more responsible, or more appropriate than AI-generated text. The challenge is that the features previously associated with human generation—natural flow, appropriate hedging, warmth, even minor errors—are now reproducible at scale.

The consequence is a strange doubling of effort. Humans must now demonstrate humanness not simply by being themselves but by being demonstrably themselves, in ways that automated systems can verify. This is a new kind of identity work, and it is not equally distributed [23]. Some people are naturally identifiable as humans by the standards that current detection systems use. Others—whose writing style, language background, or communication mode does not match the expected signature of humanness—may find themselves persistently misclassified [24].

This raises what is perhaps the deepest question the ITT

opens up: is humanness about what we can do, about what AI can do to or for us, or about how we are grounded in a physical, social, and biographical reality that machines, whatever their capabilities, do not share in the same way? The floor-based ITT we have proposed leans toward the third possibility, but likewise it holds the question open, because the more honest position is that we do not yet know how far machine simulation of human behaviour can go. More pressingly, we do not know whether there is anything that would remain distinctively human if the simulation became complete—some irreducible residue that machines cannot reach—or whether the very question of what it means to be human would dissolve if everything human could be imitated well enough.

CONCLUSION

The TT was designed to ask whether a machine could rise to the level of a human. The ITT asks whether a human can demonstrate that they have not fallen below the level of a machine. But now that there is significant overlap between human abilities and machine abilities, both questions now face significant difficulties, and for the same structural reason.

Every phase of CAPTCHA history imposed a ceiling-based structure on human authentication, asking humans to excel at something in order to qualify as human. The floor-based alternative we have proposed is more honest about what humanness actually is: not proximity to a performance ceiling, but occupancy of a range that includes inconsistency, limitation, biographical accumulation, temporal depth, social responsibility, and productive imperfection.

The four criteria we have examined also reveal a deeper problem. Each one identifies something genuinely distinctive about human existence, yet each one, on inspection, cannot be operationalised as a test without becoming a specification for simulation. Embodied situatedness, temporal depth, social responsibility, productive imperfection: these are not just difficult to test for; once they are named as sufficient to pass a test, they can be increasingly engineered as targets for machines to simulate.

This is the ITT's central paradox. The more rigorously we attempt to specify what makes humans human—by spelling out markers of humanness such as cognitive limitations, linguistic habits, emotional responses, and so on—the more detailed a blueprint we provide for machines and their designers. In time, AI systems could be designed to mimic the very traits that make us distinctively human (even though mimicking these traits is not equivalent to possessing them). This makes those systems increasingly able to pass a traditional TT. Ironically, by clarifying the qualities that separate us from machines, we equip machines with the opportunity to blur that boundary, producing agents that can simulate humanness with ever greater sophistication. In this way, the ITT does not simply test humans against machines; it also catalyses a feedback loop in which our understanding of humanness evolves and is continually redefined as we see machines become ever more able to behave like us.

What does this mean for the Turing Test itself? There is still "life" in the TT, but not as a one-off benchmark to be passed or failed. Its enduring value lies in how TT-like and ITT-like setups expose the structural instability of any attempt to fix a boundary between human and machine. The question they keep reopening—what a human being actually is—may be one we have a stake in not answering too precisely. The machines, in pressing us towards an answer, may be doing us a service we did not ask for, and one we should be cautious about accepting.

REFERENCES

[1] Graham Oppy and David Dowe. The Turing test. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Fall 2021 edition, 2021.

[2] Big Think. The Turing test: AI still hasn't passed the "imitation game". Big Think (online), March 2022.

[3] Melanie Mitchell. The Turing test and our shifting conceptions of intelligence. *Science*, 385(6710):eadq9356, 2024.

[4] Alan M. Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950.

[5] Robert M. French. The Turing test: The first 50 years. *Trends in Cognitive Sciences*, 4(3):115–122, 2000.

[6] Scott Berinato. Attack of the bots. Wired (online), 2006.

[7] Luis von Ahn, Benjamin Maurer, Colin McMillen, David Abraham, and Manuel Blum. reCAPTCHA: Human-based character recognition via web security measures. *Science*, 321(5895):1465–1468, 2008.

[8] Mohamed Akrout Ismail Akrout, Amal Feriani. Hacking Google reCAPTCHA v3 using reinforcement learning. *arXiv preprint arXiv:1903.01003*, 2019.

[9] Tom Zahavy. LLMs can't jump. *PhilSci-Archive*, 2026. Preprint, PhilSci-Archive, Archive Number 28024.

[10] Charles Sanders Peirce. Pragmatism as a principle and method of right thinking. In Patricia Ann Turrisi, editor, *The 1903 Harvard Lectures on Pragmatism*. State University of New York Press, Albany, NY, 1997.

[11] Mario Krenn, Robert Pollice, Si Yue Guo, Matteo Aldeghi, Alba Cervera-Lierta, Pascal Friederich, Gabriel dos Passos Gomes, et al. On scientific understanding with artificial intelligence. *Nature Reviews Physics*, 4:761–769, 2022.

[12] William C. Wimsatt. *Re-Engineering Philosophy for Limited Beings: Piecewise Approximations to Reality*. Harvard University Press, Cambridge, MA, 2007.

[13] Hubert L. Dreyfus. *What Computers Can't Do: A Critique of Artificial Reason*. Harper and Row, New York, 1972.

[14] Maurice Merleau-Ponty. *Phenomenology of Perception*. Routledge, London, 2012. Originally published as *Phénoménologie de la perception*. Gallimard, Paris, 1945.

[15] Daniel L. Schacter. *The Seven Sins of Memory: How the Mind Forgets and Remembers*. Houghton Mifflin, Boston, MA, 2001.

[16] Elizabeth F. Loftus and John C. Palmer. Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior*, 13(5):585–589, 1974.

[17] Elizabeth S. Parker, Larry Cahill, and James L. McGaugh. A case of unusual autobiographical remembering. *Neurocase*, 12(1):35–49, 2006.

[18] Muireann Irish, Brian A. Lawlor, Robert F. Coen, and Shane M. O'Mara. Everyday episodic memory in amnestic mild cognitive impairment: A preliminary investigation. *BMC Neuroscience*, 12:80, 2011.

[19] Peter F. Strawson. Freedom and resentment. *Proceedings of the British Academy*, 48:1–25, 1962.

[20] Daniel W. Tigard. Artificial moral responsibility: How we can and cannot hold machines responsible. *Cambridge Quarterly of Healthcare Ethics*, 30(3):435–447, 2021.

[21] Elliot Aronson, Ben Willerman, and Joanne Floyd. The effect of a pratfall on increasing interpersonal attractiveness. *Psychonomic Science*, 4(6):227–228, 1966.

[22] Daniel Wu. Some people think AI writing has a tell — the em dash. writers disagree. The Washington Post (online), April 2025.

[23] Virginia Eubanks. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press, New York, 2018.

[24] Nina Markl. Language variation and algorithmic bias: Understanding algorithmic bias in British English automatic speech recognition. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*, pages 521–534, 2022.