Cognitive Systems Research 99 (2026) 101502

Contents lists available at ScienceDirect

Cognitive Systems Research

journal homepage: www.elsevier.com/locate/cogsys

Turing's imitation game as an idealised experiment☆

Rocco Gaudenzi *, Massimiliano Badino

Department of Human Sciences, University of Verona, Verona, Italy

ARTICLE INFO

Keywords:
Imitation game
Turing test
Alan Turing
Thought experiments
Empirical experiments
Idealised experiments
Statistical hypothesis test

ABSTRACT

The literature on Turing's Imitation Game and the "Turing Test" has largely been dominated by debates about the nature and definition of intelligence, or by attempts to classify the proposal either as a ready-to-run empirical experiment or as a philosophical thought experiment. This paper argues that both readings miss a crucial structural feature. Focusing on the 1950 gender-based Imitation Game, we reconstruct it as an idealised experiment with the design-level structure of a Fisherian hypothesis test. Unlike later jury-based formats, the Imitation Game incorporates a human–human control condition and a symmetric decision task that neutralises conservative anti-machine bias. Indistinguishability is measured indirectly, via the interrogator's strategic behaviour relative to a baseline. In this reconstruction, the game does not define intelligence but probes a more specific dispositional ability in linguistic interaction. By shifting attention from definitional debates to experimental structure, the paper offers a novel methodological interpretation of Turing's proposal and its limitations. We argue that this reconstructed framework provides a rigorous methodological template that contemporary AI researchers should adopt—or explicitly depart from—when designing behavioural evaluations of machine intelligence.

1. Introduction

Since its proposal in 1950, Alan Turing's Imitation Game (IG) has occupied a peculiar and controversial position at the intersection of philosophy, artificial intelligence, and the methodology of science. Introduced in "Computing Machinery and Intelligence" as a replacement for the ill-defined question "Can machines think?", the IG has been widely discussed, criticised, reinterpreted, and often conflated with what is now commonly referred to as the "Turing Test." Arguably, the most striking feature of the IG is its experimental or quasi-experimental structure. However, after many decades and despite a vast amount of ink spilled on the issue, there still remains little agreement on the nature of this experiment.

Much of the philosophical discussion has focused on the conceptual implications of the IG, treating it primarily as a device for probing the nature of intelligence or for testing our intuitions about what should count as thinking. On this reading, the IG is often assimilated to the tradition of philosophical thought experiments, comparable to classical cases such as Galileo's falling bodies or, in more recent philosophy of mind, Searle's Chinese Room. This approach is represented in different forms by authors who interpret the IG as proposing a definition of intelligence—whether winning it is necessary, sufficient, or both for intelligence—or as offering a conceptual reformulation of the question of intelligence without genuine empirical ambitions. In this vein, Gonçalves (2023a, 2023b) has recently provided a systematic analysis of the literature, arguing that the IG should be understood as a thought experiment whose epistemic function is primarily heuristic and critical rather than test-oriented.

At the same time, a different line of interpretation has insisted that the IG has genuine experimental aspirations. Jack Copeland (2000), in particular, has argued forcefully that the IG was never intended to provide either necessary or sufficient conditions for intelligence, rejecting the widespread assumption that Turing was proposing a logical equivalence between winning the game and being intelligent. On Copeland's reading, the IG is better understood as a constrained experimental probe, whose value lies in exploring a limited domain of linguistic performance rather than settling metaphysical questions about the nature of mind.

Yet, this reading raises a further question that remains insufficiently addressed: if winning the IG is neither necessary nor sufficient for intelligence, what exactly does the game test for, and what kind of knowledge, if any, does it yield? Copeland himself suggests that the IG

☆ This article is part of a Special issue entitled: 'Machines in Thought' published in Cognitive Systems Research.

* Corresponding author.
E-mail address: rocco.gaudenzi@univr.it (R. Gaudenzi).

https://doi.org/10.1016/j.cogsys.2026.101502
Received 16 February 2026; Received in revised form 6 May 2026; Accepted 16 July 2026
Available online 18 July 2026
1389-0417/© 2026 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

measures linguistic abilities but leaves open whether these should be understood as indirect proxies for intelligence or as some more limited competence. Closely related to this issue is the question of how to interpret Turing's oft-repeated remarks that intelligence is something "emotional" and therefore that judgments of intelligence depend on the observer, a point emphasised by Proudfoot (2014) and Numerico (2005). If attributions of intelligence vary with the interrogator, then any experimental design must confront the problem of noise introduced by human judgment and must provide a way—if possible—to screen or control it.

This paper proposes a new way of addressing these interconnected issues by arguing that the 1950 version of the IG is best reconstructed neither as a pure thought experiment nor as a full-fledged empirical experiment, but as an *idealised experiment*. Idealised experiments stand halfway between thought experiments, whose epistemic value lies entirely in pondering conceptual relations, and empirical experiments, whose value lies in putting the world to the test. Reconstructing the IG as an idealised experiment allows us to explain its philosophical generativity, but also why it has been viewed at times as a mere conceptual device, at times as a ready-made empirical protocol.

We submit that the IG has the structure of a statistical hypothesis test whose target is not adjudicating intelligence in general but a specific dispositional capacity: the capacity to sustain human-like linguistic performance under specific pragmatic and conversational constraints. Under these conditions, the machine's task is to act indistinguishably from a human being, which, in terms of statistical inference, amounts to forcing the human interrogator to resort to chance in making their evaluation. This framework clarifies what falls within the IG's epistemic scope, why this scope is necessarily limited, and why later attempts to implement "Turing Tests" in practice—most notably the Loebner Prize competitions—have suffered from severe methodological flaws when the idealised structure of the original game was ignored.

The paper is organised as follows. Section 2 reviews the different versions of the IG and situates the present proposal within the existing literature, drawing in particular on Gonçalves's categorisation of interpretations, the analysis of the game's variants by various authors, and Copeland's critique of definitional readings. Section 3 introduces the notion of idealised experiment and argues that the 1950 IG fits this category, both historically and methodologically. Section 4 develops the statistical hypothesis-testing structure of the game and clarifies the dispositional capacity it is designed to probe, filling in some of the missing implementation details that an empirical realisation would require. Section 5 examines the limits of the IG, focusing on the problem of noise introduced by human interrogators and drawing lessons from real-world attempts to conduct Turing-like tests. The concluding section summarises the implications of this reconstruction for the philosophy of artificial intelligence and for the broader understanding of experimentation in contexts where human judgment itself is part of the measurement apparatus.

2. The imitation game in the literature: corpus, interpretations, and target properties

This section situates the IG within the existing literature by examining three closely related issues that have shaped its interpretation. We begin by delimiting Turing's corpus and reviewing how different authors distinguished—or failed to distinguish—between the various game-like procedures involving human and machine interaction described in his writings. We then survey the main interpretive traditions concerning the epistemic role of the IG, with particular attention to the contrast, systematised by Gonçalves, between readings that regard it as an inadequate experiment and those that defend it as a meaningful test for artificial intelligence. Finally, we consider what property the IG has been taken to test in the empirical tradition, what it is supposed to measure, and what conclusions the game is thought to license with respect to machine intelligence. Taken together, these strands of the literature reveal persistent tensions concerning the experimental status of the IG, tensions that motivate the interpretive framework developed in the following sections.

2.1. Turing's corpus and the status of the IG

Any assessment of the experimental status of the IG must begin with an analysis of Turing's work. Scholars disagree more about the relationship between Turing's various game-like procedures involving humans and machines than they do about interpretations of a single, well-defined test. Therefore, before turning to interpretive disputes, it is necessary to outline the core versions that are universally accepted as being present in Turing's writings and broadcasts.

A first version appears in *Intelligent Machinery* (1948), where Turing describes an explicitly experimental situation involving chess. The setup involves three agents. Two of them, A and C, are assumed to be chess players—Turing stresses that they should not be excellent players, but rather average or even poor ones. The third agent, B, operates what Turing calls a "paper machine": a person who manually simulates a chess-playing algorithm by following rules written on paper. The interrogator-player C plays chess against either A or B, knowing that one of the two possible opponents is the simulated machine but not knowing which one is currently being faced. The task is to determine, on the basis of play alone, whether the opponent is the human or the machine operator. In this version, the decision problem is asymmetric: the interrogator must single out the machine from among possible opponents, and the presence of a machine is explicitly part of the epistemic background. As in the later 1952 formulation, the structure of the task permits a conservative strategy under uncertainty, since identifying the machine is the central objective.

The second, and for present purposes central, version is introduced in *Computing Machinery and Intelligence* (1950). Here Turing replaces chess with unrestricted linguistic interaction and introduces what he calls the "imitation game". In its canonical formulation, the game involves three participants: a man, a woman, and a human interrogator who communicates with them in writing and must determine who is who. The man attempts to deceive the interrogator by pretending to be the woman, while the woman answers truthfully. Turing then asks what will happen "when a machine takes the part of A [the man] in this game," and whether the interrogator will "decide wrongly as often" as in the human–human case (Turing, 1950, p. 434). A crucial feature of this formulation is the comparative structure it establishes between two conditions. In the control condition, the interrogator assigns the labels "man" and "woman" to two human players; in the test condition, the man is replaced by a machine, and the interrogator's task becomes to assign the labels "machine" and "woman." Unlike in the 1948 and 1952 versions, where the task is to single out the machine, in both control and test conditions the interrogator must distribute two labels between two respondents. This symmetry of the decision space is methodologically significant: since both labels must be assigned and no category can be chosen as a safer default option, the most rational and prudential strategy in uncertain conditions—that is, the course that minimises expected error in the absence of any fact of the matter distinguishing human and machine—is to assign labels randomly.

A third version appears shortly thereafter, most explicitly in Turing's 1952 BBC radio broadcast. In this later presentation, the gender element drops out, and the setup is often described as a direct conversational test in which an interrogator—or a jury—tries to identify the machines among the interlocutors. As in the 1948 version, Turing emphasises that interrogators should be "average" and not experts on machines, and that judgments will inevitably be fallible. This version, now commonly known as the "Turing Test", retains the conversational format but does not include an explicit human–human control condition. Moreover, the decision space is no longer symmetric: the interrogator knows that some of the interlocutors are machines and must identify which ones. In such a setting, uncertainty can rationally generate a prudential bias. Mistaking

a machine for a human may be perceived as epistemically embarrassing, whereas mistaking a human for a machine is typically less costly. Under these conditions, an interrogator may adopt a conservative strategy—"when in doubt, say machine"—thereby skewing results in a way that is structurally unavailable in the gender-based imitation game. In Section 5, we return to the broader question of how such decision-theoretic considerations and other sources of bias affect the interpretation of Turing-style experiments.

These three versions—the chess-based imitation game (1948), the gender-based imitation game (1950), and the simplified conversational test (1952)—form the core of Turing's corpus on imitation and machine intelligence (Proudfoot, 2022).

Within this corpus, an important refinement has been proposed by Sterrett (2000), who argues that even the 1950 paper contains two distinct formulations of the imitation game. On her reading, Turing first introduces what she calls the 'Original Imitation Game', in which success is defined in terms of matching the human–human error rate. Later in the same paper, Turing shifts to a formulation in which the interrogator's task is more naturally read as identifying which participant is the machine, where the interrogator plays one to one with the machine. Sterrett refers to the latter as the 'Standard Turing Test' and claims that much of the confusion in the literature arises from conflating the two. When we refer to the IG, we are referring to Sterrett's Original Imitation Game.

Focusing on this version enables us to highlight the key interpretative disagreement that has influenced subsequent literature. Broadly speaking, authors have disagreed over whether the IG (or the Turing Test more generally) should be considered a meaningful experiment in the field of artificial intelligence, or whether it is so poorly defined and flawed that it cannot support serious experimentation. Gonçalves (2023a, 2023b) characterised this opposition as the 'Turing test dilemma'. One side of the debate argues that the test is too ambiguous, biased or susceptible to manipulation to count as a scientific experiment, while the other side defends it as a powerful and general test that is too valuable to abandon, despite the practical difficulties of implementing it.

As we argue in the remainder of this section, this dilemma arises in part from treating Turing's proposal as if it must be either a thought experiment or a fully specified empirical experiment. The gender-based imitation game of 1950 fits neither category neatly. It specifies the design of an experiment while leaving crucial implementation details open. Concentrating on this version allows us to assess, in a precise way, what kind of experiment Turing was proposing, what sort of capacity it is designed to test, and why its apparent shortcomings are better understood as features of an idealised experimental design rather than as fatal flaws.

2.2. What does the IG test?

A third recurring theme in the literature concerns what property, process, or fact of the matter, if any, the IG is supposed to test and what follows from success or failure in the game. Different answers are offered depending on whether the IG is interpreted conceptually or empirically.

In the empirically oriented tradition, some authors have suggested that the IG tests the ability to think like a human (Moor, 1976; Dennett, 2006), or intelligence tout court, sometimes construed as a dispositional property (Block, 1981). Others adopt a more behaviourist stance, treating the test as measuring the ability to behave like a human or to disguise oneself successfully as one (French, 1990). Copeland (2000), by contrast, offers a more restrained interpretation, arguing that the IG primarily probes linguistic abilities rather than intelligence in any robust metaphysical sense. Crucially, he shows that passing the test cannot reasonably be considered a necessary or sufficient condition of intelligence. This challenges the interpretations of the test's definition proposed by Osherson and Lasnik (1990) and Hodges (2014).

These disagreements are closely tied to differing views about the metaphysical status of intelligence itself—further complicated by the absence of a univocal definition of intelligence. Some authors treat intelligence as an intrinsic property of systems, others as a relational property defined by interaction, and still others as observer-dependent—see Section 4.1 for details. Diane Proudfoot (2014), the main supporter of the latter, argues that mere behaviour—understood as a property of the object under consideration in the IG—does not suffice for intelligence or thinking, highlighting the role of the observer's stance in attribution.

In this paper, we adopt Copeland's modest view of the scope of the IG. We agree that the IG does not test intelligence directly, but rather a more specific set of linguistic abilities. However, we also argue that these abilities are best understood as dispositional, manifesting only under specific interactional and pragmatic conditions, and unable to be identified by inspection alone. Linguistic performance in the IG may serve as an indicator of broader capacities, such as generalisation and learning, which are widely regarded as necessary conditions for intelligent behaviour when that is defined as the capacity to select and execute actions that further situational goals under conditions of uncertainty and with limited energy, time, and computational resources (Abbate & Gaudenzi, 2025).

3. The IG as an idealised experiment

3.1. Idealised experiments: between thought experiments and empirical experiments

In recent scholarship, especially in the work of Sophie Roux (2011), a third category has emerged alongside the traditional distinction between thought experiments (TEs) and empirical experiments (EEs): the category of idealised experiments (IEs). Idealised experiments occupy a conceptual middle ground: Their design is specified in sufficient detail to articulate clear experimental roles, variables, and an inferential structure, yet they do not supply the full set of practical or material specifications required for immediate implementation. In this sense, IEs are best characterised as potential experiments: they are neither mere heuristic fictions nor fully instantiated laboratory protocols.

Thought experiments, from Galileo's falling stones to Putnam's Twin Earth, operate entirely in the imagination. Their epistemic aims are conceptual: They are designed to probe intuitions, expose hidden assumptions, test the coherence of a theory, or reveal contradictions in an existing conceptual framework. Their success does not depend on empirical execution, and empirical execution, when possible, would not normally add further epistemic value. A TE requires only logical and physical coherence; it does not require parameters for replication, control of sources of noise, or instructions for statistically interpreting outcomes. IEs, by contrast, specify their structure with empirical execution in mind. They provide a defined set of roles, such as experimental subjects, interrogators and experimenters; articulate variables and expected outcomes; and supply a methodological rationale for inference. However, they are incomplete by design. For example, an IE does not specify every operational detail that an EE requires, such as the number of trials, the method for selecting participants, the procedures for randomisation, or precise evaluation criteria. An IE typically focuses only on factors that are considered causally relevant to the phenomenon under study.

Empirical experiments are materially realised. They require the instantiation of the protocol, the calibration of instruments, the explicit definition of noise sources and strategies for mitigating them, and the statistical treatment of data collected in execution. An idealised experiment, by contrast, precedes empirical execution. It provides the conceptual and methodological scaffolding for an empirical experiment while leaving certain details open to future specification. In this way, IEs are oriented toward empirical investigation without being identical to it.

The epistemic roles of idealised experiments reflect this dual character. Like thought experiments, they have heuristic and critical

functions: they clarify hypotheses, illuminate conceptual commitments, and explore the inferential consequences of a given setup. But unlike thought experiments, they also play a normative methodological role. They outline how empirical inquiry should proceed, set constraints on valid inference, identify conditions necessary for reliable measurement, and provide an experimental template ready for empirical completion. Their epistemic force thus derives both from the contemplation of the design itself and from the prospect of empirical execution. The design conveys methodological insight, while its potential empirical realisation yields data according to the structure the IE prescribes.

Viewed through this lens, Turing's 1950 IG fits squarely within the category of idealised experiments. Its structure is explicitly experimental: it assigns defined roles to participants, incorporates controlled substitution, and specifies an outcome measure in terms of the frequency of misidentification. What it lacks are the contingent implementation details needed to transform it into a fully empirical experiment, such as the number of interrogators, parameters for sample selection, or the duration of the game. These are precisely the elements an experimenter, rather than a theoretician, would provide.

Moreover, the design is empirically directed. Turing repeatedly stresses the use of statistics and control conditions, which are hallmarks of empirical experimental design, yet he leaves the operational numbers and procedural refinements unspecified. As a result, the IG occupies a position that naturally invites dual interpretation: those who emphasise its conceptual provocations naturally classify it as a thought experiment, while those who emphasise its methodological scaffolding view it as an empirical protocol. In reality, it is neither exclusively nor primarily one or the other; it is an idealised experiment whose structure embodies the logic of a scientific test while performing, at the textual level, many of the functions of a thought experiment.

3.2. Why IG has the structure of a hypothesis test

If the IG is best understood as an idealised experiment, the next step is to clarify what kind of experiment it is. We argue that the 1950 IG has the design-level structure of a Fisherian hypothesis test, even though Turing did not articulate it in formal statistical terminology. To make this claim precise, it is essential to distinguish between elements that belong to the experimental design—which determine the inferential logic of the test—and elements that belong to its implementation. Following the methodology adopted in Fisher and Prance (1974), design-level elements include the specification of trials, outcome variables, control and test conditions, and the formulation of the null hypothesis. Implementation-level elements include sample size, duration, sampling procedures, and statistical significance thresholds. Idealised experiments characteristically specify the former while leaving the latter underdetermined. As we show in the following sections, Turing's presentation of IG fully specifies the design-level elements of a hypothesis test, while leaving implementation details to future experimenters. The fact that Turing did not himself implement the 1950 version does not undermine its experimental status; it marks it as an idealised experiment. There is at least one previous evidence suggesting that presenting idealised experimental designs is consistent with Turing's own practice: he explicitly reports having performed the chess-based version discussed in 1948, describing it as “a rather idealized form of an experiment I have actually done” (Turing, 1948, p. 431).

3.2.1. Control and test conditions

To understand the hypothesis-testing structure of IG, we begin with the two conditions it establishes and the benchmark they provide. IG is organised around a control condition and a test condition. In the control condition, a man and a woman respond in writing to questions posed by a human interrogator, whose task is to determine who is who. Because men and women are assumed, on average, to be equally capable of deception and truth-telling, no systematic basis for discrimination is available. Under this assumption of human symmetry, the rational course of action is to randomise, yielding an expected error rate of 50%. This baseline serves as the reference point for evaluating performance in the test condition.

In the test condition, the man is replaced by a machine, and the experimental question becomes whether, across repeated trials, the interrogator performs significantly better than in the control condition. What the experimenter tracks is not the machine's internal capacity directly, but the interrogator's shift in strategy. If performance remains at baseline, this indicates continued reliance on random choice and thus indistinguishability. A systematic deviation from the baseline, by contrast, signals that the interrogator has access to performance differences sufficient to sustain a non-random strategy. Indistinguishability is therefore measured indirectly, through the stability or breakdown of chance-level performance across conditions.[1]

3.2.2. The hypotheses embedded in the IG

Given this experimental structure, IG naturally embeds the two hypotheses characteristic of a Fisherian hypothesis test. The null hypothesis expresses experimental indifference: performance in the test condition does not differ from the baseline established in the control condition. Under the null, no discriminating fact of the matter is available to the interrogator beyond what was present in the control setting. The alternative hypothesis states that performance in the test condition deviates systematically from this baseline, indicating that the machine is distinguishable from the human respondent.

This formulation reflects the orientation adopted by Fisher in the famous tea-tasting lady experiment, in which the null represents the absence of a detectable effect. Importantly, it keeps the inferential structure distinct from Turing's empirical expectations. Although Turing predicted that early machines would be distinguishable, such expectations concern anticipated outcomes, not the logical architecture of the test itself. The design remains neutral with respect to results: both indistinguishability and distinguishability are assessed relative to the same experimentally established baseline.

3.2.3. Outcome measure and repeatability

The outcome measured by the IG is binary: either the interrogator's judgment is correct, or it is not. This outcome is discrete, repeatable, and amenable to statistical analysis. Each completed game constitutes a trial yielding a binary observation. Across multiple trials and interrogators, the experimenter can estimate the frequency with which the machine is misidentified relative to the human baseline. This structure is precisely that of a hypothesis test: the experimental question is whether the observed proportion of correct identifications falls within the distribution expected under the null hypothesis.

3.2.4. Idealisation and the role of implementation details

Taken together, these elements show that the 1950 IG fully specifies

[1] This reconstruction highlights an idealising assumption that was implicit in Turing's design: that the agents involved behave rationally. For the interrogator, this means adopting a random strategy when there is no evidence to discriminate, since this minimises expected error in a symmetric decision space. For the players, rationality means avoiding extraneous tricks that would merely trigger the interrogator's biases rather than revealing genuine differences in performance. These are structural assumptions that render the hypothesis-testing framework coherent, not empirical claims about actual human behaviour. In Section 5, we return to the limits of these idealisations. More generally, we draw the attention of the reader to the fact that the outcome of the game depends not only on its formal structure, but also on the information provided to the interrogator. While the experimenter operates within a hypothesis-testing framework (one might imagine Turing as the first to be in this role), the interrogator faces a decision-theoretic problem: how to allocate labels under uncertainty, given the rules of the game. The interrogator's strategic options, and therefore the interpretation of the resulting statistics, are directly shaped by what they are told about the identities of the players, the possibility of machine participation, and who may be lying. The simplest configuration that minimises bias and maximises symmetry between the control and test conditions appears to be one in which the interrogator is informed of the relevant label categories in each condition, and that one of the two players is lying.

the design of a Fisherian hypothesis test while remaining underspecified at the level of implementation. This corresponds to IEs in Roux's terminology: one that yields methodological insight through its design while allowing multiple empirical realisations.

Turing's own quantitative prediction illustrates this point. When he suggests that machines with such-and-such features (he speaks in particular of storage capacity) might one day "play the imitation game so well that an average interrogator will not have more than a 70 percent chance of making the right identification after five minutes of questioning" (Turing, 1950, p. 442), he advances a possible future outcome and proposes a duration for the game. These remarks concern implementation, not design. They show how the idealised experiment might be instantiated under particular assumptions, without fixing the parameters required for a fully operational protocol. In Section 4, we address this task by proposing how the missing implementation details might be filled in to best measure the dispositional property that, in our view, IG is designed to test.

4. Intelligence as a dispositional capacity

If IG is best reconstructed as an idealised experiment—one that specifies a rigorous experimental structure while leaving operational details for later implementation—we must now ask a more fundamental question: What exactly does such an experiment test for? Since, as Copeland (2000) convincingly argues, the IG cannot provide a necessary or sufficient condition for intelligence, its explanatory value must lie elsewhere. Our claim is that IG is designed to test for a dispositional capacity, namely the capacity of a machine to produce human-like linguistic performance under conditions of strategic interaction and epistemic neutrality. In this section, we clarify what kind of property this is, why it is dispositional rather than intrinsic or observer-dependent, and how this interpretation makes sense of both the experimental logic and the historical context of Turing's proposal.

4.1. Dispositional, intrinsic, relational, and observer-dependent properties

Properties fall into broad metaphysical categories. Intrinsic properties are possessed independently of context (e.g., mass, shape) (Lewis, 1983; Weatherson & Marshall, 2023); relational properties depend on a relation to something else (e.g., being taller than one's sibling); response- or observer-dependent properties depend partly on an observer's stance or judgment (e.g., beauty, humour) (Johnston (1989), Pettit (1991)). By contrast, dispositional properties—fragility, solubility, courage, irritability—manifest only under specific conditions, when "triggered" by an appropriate setup. They cannot be inspected directly; their reality is inferred through performance given a suitable test environment (Bird, 2007; Choi & Fara, 2021; Mumford, 1998).

A disposition is always defined by the interplay between an underlying capacity, a triggering condition, and a characteristic manifestation. The IG fits precisely this structure. It establishes a triggering condition—substitution of a machine into a constrained dialogue—which enables the manifestation (or failure) of a capacity for human-level linguistic performance. Unlike intrinsic properties (e.g., neural architectures), a capacity for strategically navigating natural language emerges only in interaction. And unlike observer-dependent properties, this capacity is not created by the interrogator's judgment; rather, the interrogator serves as a probe, a measurement instrument with noise that must be controlled for in the design.

This dispositional reading resolves an interpretive tension in the literature. Scholars such as Proudfoot (2022, 2014)) and Numerico (2005) emphasise that Turing framed intelligence as an "emotional concept", manifesting thus through social attribution and belonging to the category of observer-dependent concepts. What gives rise to interpretation is Turing's passage stating that "the extent to which we regard something as behaving intelligently is determined as much by our state of mind and training as by the properties of the object under consideration" (Turing, 1948, p. 431). That is correct—but the fact that intelligence manifests through observer response does not make it observer-dependent. Fragility manifests when someone drops a glass, but it does not depend on who drops it. In fact, we regard Turing's statement as manifesting awareness that the experimenter's task is to screen from that part of the judgment that is observer-dependent and, in this sense, "emotional". The distinction between kinds of properties becomes crucial for understanding what IG can and cannot test.

4.2. The dispositional capacity measured by the IG and its limits

Any assessment of what the IG measures presupposes at least a minimal clarification of what is meant by intelligence. Given the plurality of definitions in the literature, no attempt will be made here to exhaust the concept. Instead, we adopt a working definition proposed by Wang (2019) and refined by Abbate and Gaudenzi (2025), according to which intelligence consists in the capacity to behave relevantly under conditions of uncertainty and limited information. In this view, intelligent behaviour is not defined by the possession of specific internal mechanisms but by the ability to pursue goals effectively in environments characterised by ill-defined problems, incomplete knowledge, and an open-ended space of possible actions.

On this account, two core abilities are minimally required for intelligence: learning, understood as the capacity to update behaviour on the basis of experience, and generalisation, understood as the capacity to apply acquired competence to novel or unforeseen situations. Generalisation itself admits degrees. Following Chollet (2019) and Abbate and Gaudenzi (2025), it is useful to distinguish between local generalisation, which concerns adaptation within a narrow and well-defined task space; broad generalisation, which concerns adaptation across a wide class of related tasks under uncertainty; and extreme generalisation, which characterises the human ability to transfer competence across heterogeneous domains of knowledge, perception, action, and social practice.

In this context, the dispositional capacity tested by the IG is not equivalent to intelligence in a general sense. Rather, it is a limited ability: the capacity to sustain human-like linguistic interaction under conditions of epistemic neutrality and strategic deception, such that ordinary human interrogators cannot distinguish the machine from a human interlocutor above chance levels. In this sense, Copeland (2000) is right to insist that the IG neither defines intelligence nor provides necessary or sufficient conditions for it.

Nevertheless, that capacity is not trivial. Linguistic interaction in the game does not take place in a narrowly specified task environment. It unfolds in a setting characterised by implicit goals, shifting conversational norms, underdetermined tasks, and contextual uncertainty. Successful participation therefore requires the ability to respond appropriately to novel prompts, manage conversational implicatures (Grice 1975), track presuppositions (Stalnaker 1977), and align with interlocutors' expectations (Austin 1975). These are the features that distinguish broad generalisation from narrow, task-bound competence (Abbate & Gaudenzi, 2025). For this reason, the dispositional capacity tested by the IG can be understood as a proxy for broad generalisation in the linguistic domain. Like many scientific proxies, it captures a core dimension of the target phenomenon without exhausting it. Broad linguistic generalisation, as understood here, is plausibly a necessary condition for human-level intelligence, but it is not sufficient for it. The IG thus tests for a capacity that is epistemically relevant to intelligence without collapsing into a definition of intelligence itself.

Once this point is made explicit, the limits of the IG follow directly. First, the game does not test intelligence as an intrinsic or internal property. It abstracts entirely from internal mechanisms and cannot assess planning, causal modelling, symbolic reasoning, or representation-building except insofar as these capacities manifest themselves in linguistic behaviour. Second, the game does not test extreme generalisation. Capacities that involve transferring competence

across non-linguistic domains—such as sensorimotor coordination, perceptual discrimination, emotional regulation, or practical know-how—lie outside its scope. Third, success in the game cannot be taken as either a necessary or a sufficient condition for intelligence. It probes a particular manifestation of intelligent behaviour, not its metaphysical essence. Finally, the IG cannot test intelligence independently of its anthropocentric and interrogator-mediated framing. The benchmark of success is human linguistic performance, and the measurement apparatus consists of human interrogators whose judgments inevitably introduce noise, bias, and normative expectations. These features are not accidental defects but structural consequences of the property the game is designed to measure.

Taken together, these limitations do not undermine the scientific significance of the IG. Rather, they delimit its epistemic reach. Properly understood, the IG is an idealised experiment that tests whether a system exhibits a form of broad generalisation in linguistic interaction—a dispositional capacity that is plausibly necessary for, but far from constitutive of, human intelligence.

4.3. Filling in the missing experimental details

As it happens often in statistical experiments, the implementation details must be supplied by the experimenter in light of two competing constraints: on the one hand, the need to minimise statistical distortion and sources of error given the dispositional property the game aims to test; on the other, the contingent limits imposed by available resources, time, and participants. In what follows, we do not aim to exhaust the space of possible implementations but to illustrate the kinds of considerations that arise once the IG is treated as a genuine experimental design.

For the sake of argument, suppose that the experimenter chooses to model the outcomes of the IG using a binomial distribution ($B(n, p)$), where each completed game of the total number of games (n) yields a binary outcome (correct or incorrect identification) and where p represents the probability of correct identification under the null hypothesis. On the assumption of human symmetry in the control condition, we expect $p = 0.5$. To justify the use of a binomial model, however, the experimenter must ensure that trials are approximately independent, which immediately raises questions about the reuse of participants across games.

One option is to replace both the pair of respondents and the interrogator at each trial. This maximises independence across trials but may be impractical given limited resources. A second option is to hold fixed a pair, or a small number of pairs, of respondents while rotating interrogators across games. This has the advantage of allowing the experimenter to select respondent pairs for which the control condition empirically approximates the assumed baseline of $p = 0.5$. If some pairs exhibit systematic asymmetries—leading to baseline probabilities that differ from chance—this need not invalidate the experiment. Rather, the estimated baseline probability can be incorporated directly into the binomial model by adjusting the value of p accordingly when evaluating performance in the test condition. In this way, departures from ideal symmetry can be treated as measurable features of the control rather than as sources of uncontrolled noise.

A further crucial implementation parameter concerns the duration of each game. The proportion of games won by the interrogator in the test condition depends at least partly on how long the interaction lasts. Longer exchanges provide more behavioural data, typically reducing uncertainty and increasing the likelihood of detection. Shorter exchanges, by contrast, may amplify fluctuation across trials. There are at least three principled ways to address this time dependence.

One approach is to define indistinguishability in a way that approximates ordinary human–human interaction by making the duration sufficiently long that additional time no longer significantly affects performance. In this approach, duration is chosen so that the interrogator's judgment stabilises, and time effectively drops out as a relevant variable. A second approach is to fix the duration arbitrarily and define (in)distinguishability relative to that fixed temporal constraint. The experiment then asks whether the machine can induce chance-level performance given a specified interaction length. A third approach is to treat duration itself as an experimental parameter and to study how the interrogator's success rate varies as a function of time. This may reveal thresholds or regimes in which performance changes qualitatively, allowing the experimenter to motivate one of the previous strategies.

Turing's own proposal aligns most closely with the second approach. When he predicts that, within fifty years, machines will "play the imitation game so well that an average interrogator will not have more than a 70 per cent chance of making the right identification after five minutes of questioning" (Turing, 1950, p. 442), he fixes a specific duration and associates it with an expected outcome. This move does not complete the experimental protocol, but it illustrates how an idealised design can be partially instantiated by selecting concrete values for implementation parameters.

These examples are sufficient to make the general point. Once the IG is understood as an idealised experiment, questions about sampling, independence, baseline probabilities, and duration are no longer objections to its experimental status but natural extensions of its design. Different choices along these dimensions correspond to different empirical realisations of the same underlying experiment, all aimed at measuring—under controlled and statistically interpretable conditions—the dispositional capacity for human-like linguistic performance that the IG is designed to probe.

4.4. Historical and methodological implications of our interpretation

Understanding the IG as an IE testing for a dispositional capacity has several important implications for both methodology and historical interpretation. First, it dissolves the Turing Test Dilemma (as described by Gonçalves): the test has both conceptual and empirical dimensions because IEs always do. It is neither a flawed experiment (negative horn) nor an entirely implementable one (positive horn), but an idealised experiment, designed to be both thought-provoking and empirically oriented. Second, it explains why the IG has been misunderstood as a TE: its conceptual sharpness naturally lends itself to philosophical exploration even in the absence of empirical data. Third, it reveals why the chess-based and jury-based versions (the latter closest to the standard Turing Tests) fail as experiments—not because they lack philosophical interest, but because their design introduces sources of bias or noise (related to the prosecutorial framing and the absence of a control condition) that prevent successful measurement of the dispositional capacity in question. Fourth, this analysis clarifies the limited sense in which linguistic performance can be used as a proxy for intelligence: aspects of intelligence such as broad generalisation manifest themselves dispositionally through linguistic interaction. Finally, it establishes the IG as a precursor to modern experimental practices in AI and cognitive science, in which behavioural indistinguishability tests—such as adversarial linguistic tasks and pragmatic reasoning benchmarks—operate using a similar logic.

The last three points have a methodological dimension that bears directly on contemporary debates about how to interpret the behavioural performance of large language models (LLMs) and, more broadly, about what kinds of experiments can support inferences about machine intelligence. Recent years have seen a proliferation of behavioural experiments on LLMs that, explicitly or implicitly, invoke Turing-style reasoning. Systems such as GPT-4 have been reported to pass empirical Turing tests (Jones & Bergen, 2024) and to display "sparks" of general intelligence across a range of behavioural probes (Bubeck et al., 2023), generating sharply divergent interpretations: some take such results as evidence of emergent human-like intelligence (sometimes vaguely labelled "general intelligence"), while others argue that they reflect sophisticated pattern-matching without underlying

understanding (Mitchell & Krakauer, 2023; Shanahan, 2024).

Our reconstruction provides analytical tools for clarifying the issues at stake in these disagreements. Following Abbate and Gaudenzi (2025), we consider extreme generalisation (discussed in Section 4.2) to be a necessary condition for human-like intelligence. IG-style linguistic behavioural testing cannot probe this capacity—only tests of cross-domain, multimodal reasoning may shed some light on it (see Bubeck et al., 2023; Maslej et al., 2024, for examples). Claims that LLMs have achieved emergent human-like intelligence on the basis of purely linguistic tests therefore overreach: the dispositional capacity such tests probe is structurally narrower than the property the claims invoke.

At the design level, the proliferation of IG-style LLM evaluations has often not been accompanied by a corresponding rigour in experimental design. The recent empirical Turing test conducted by Jones and Bergen (2024), for example, adopts a jury-based identification format, with the aforementioned limitations. This does not imply that such results are uninformative. It does suggest, however, that the IG, properly understood as an idealised experiment, offers a methodological template—control condition, symmetric decision space, baseline-relative measurement—that current behavioural probes of machine cognition would do well either to adopt or to depart from with explicit justification.

5. Critical points and limits of the IG

The preceding reconstruction clarifies the inferential structure of the IG, but it also raises important questions. First, to what extent can the design realistically screen off noise arising from the three human participants, who serve simultaneously as measurement instruments (the interrogator in particular) and potential sources of distortion? Second, what forms of bias remain even under the symmetric decision structure of the game? This section addresses these questions and evaluates the limits of IG in light of both its design logic and later attempts to implement Turing-style tests.

5.1. Noise, bias, and the human participants

A central methodological challenge in IG is that the interrogator is not a neutral measuring device but a human agent whose judgments are shaped by cognitive biases, expectations, fatigue, varying degrees of rationality, and differences in their linguistic sensitivity, strategic sophistication, and cognitive style. In statistical terms, this means the interrogator introduces both random and systematic measurement noise.

Unlike in the later standard Turing Test, the symmetry of decision costs in the gender-based IG ensures that randomisation, rather than category bias, is the rational response under uncertainty, since this minimises expected error in a symmetric decision space. This, however, depends on an idealising assumption implicit in Turing's design: that agents behave rationally[1]. For the interrogator, this means adopting a random strategy when evidence is insufficient to discriminate, rather than relying on idiosyncratic heuristics, e.g., the stereotype-driven one that may map perceived performance differences onto gender labels. For the respondents, it means avoiding extraneous tricks that merely trigger bias rather than reveal genuine performance differences.

Systematic deviation from chance performance in the control condition signals either a departure from rational decision-making or a breakdown in the assumption of human symmetry—due to the presence of actual asymmetries in the truth-telling and deception abilities of the respondent sample. While such departures cannot be entirely eliminated, the control condition serves as both baseline (see Section 4.3) and diagnostic tool for evaluating the extent to which these idealising assumptions hold in practice.

5.2. Lessons from real-world turing tests

The advantages of IG's design become clearer when contrasted with later implementations of Turing-style tests, most notably the Loebner Prize competitions (1991–2019). As documented by Epstein (2009), Saygin et al. (2000) and Shieber (1994, 2016), these implementations suffer from structural weaknesses that the IG avoids.

Most critically, practical Turing tests combined two design flaws that introduced systematic bias: they lacked human–human control conditions and adopted jury-based formats where interrogators knew one or more interlocutors were machines. This creates asymmetric decision costs that encourage conservative identification strategies under uncertainty (Hayes & Ford, 1995). Without a baseline to distinguish machine detectability from interrogator bias, such results are uninterpretable. Additionally, implementations typically suffered from poor statistical design—inconsistent evaluation metrics, subjective "humanness" scores rather than binary outcomes, and extremely small sample sizes that preclude reliable inference (Christian, 2011; Floridi et al., 2009; Shieber, 2016). Finally, practical tests often allowed extraneous cues—typing speed, formatting, latency—to influence judgments, shifting focus away from the linguistic-pragmatic capacities the IG targets (French, 2000).

These shortcomings illustrate how easily the inferential logic of the IG collapses when its key structural features—symmetric decision space, explicit control condition, and clearly defined dependent variable—are abandoned. The IG does not eliminate all noise, but it minimises structural biases and provides internal resources for diagnosing others.

6. Conclusion

By reconstructing Turing's 1950 Imitation Game as an idealised experiment, we resolve a seventy-year interpretive puzzle while clarifying what behavioural tests can and cannot tell us about machine intelligence. The IG is neither pure philosophy nor ready-made science, but a methodological template that shows how to probe intelligence through controlled interaction—testing for broad linguistic generalisation rather than intelligence itself. This framework is immediately relevant for contemporary debates in at least two ways. First, as large language models increasingly pass ad hoc "Turing tests," the distinction between linguistic competence and general intelligence becomes critical. Claims about emergent machine intelligence based on purely linguistic performance systematically overreach, mistaking an arguably necessary condition for a sufficient one.

Second, in an era where behavioural evaluation of AI systems is proliferating rapidly, Turing's idealised experiment remains surprisingly modern: Its design principles—explicit baselines, symmetric decision structures, and hypothesis-driven evaluation—offer methodological rigour that current AI assessments often lack. We invite AI researchers to reconsider it as an experimental template.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work the author(s) used Chat-GPT in order to proofread the manuscript. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

CRediT authorship contribution statement

Rocco Gaudenzi: Writing – review & editing, Writing – original draft, Conceptualization. **Massimiliano Badino:** Writing – review & editing, Conceptualization.

Funding

The authors declare no funding.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors declare no funding.

Data availability

No data was used for the research described in the article.

References

Abbate, F., & Gaudenzi, R. (2025). Ungrounding intelligence: From biological minds to artificial cognition. *Philosophical Psychology*, 1–30.
Bird, A. (2007). *Nature's metaphysics: Laws and Properties*. Oxford University Press.
Block, N. (1981). Psychologism and behaviourism. *Philosophical Review, 90*, 5–43.
Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., … Zhang, Y. (2023). *Sparks of artificial general intelligence: Early experiments with GPT-4*. arXiv preprint arXiv:2303.12712. https://doi.org/10.48550/arXiv.2303.12712.
Choi, S., & Fara, M. (2021). Dispositions. In *Stanford encyclopedia of philosophy*.
Chollet, F. (2019). *On the measure of intelligence*. arXiv preprint arXiv: 1911.01547.
Christian, B. (2011). *The most human human*. Doubleday.
Copeland, B. J. (2000). The Turing test. *Minds and Machines, 10*(4), 519–539. https://doi.org/10.1023/A:1011285919106
Dennett, D. C. (2006). *Sweet dreams: Philosophical obstacles to a science of consciousness*. MIT Press.
Epstein, R. (2009). The quest for the thinking computer. In R. Epstein, G. Roberts, & Beber (Eds.), *Parsing the Turing test*. Dordrecht: Springer.
Fisher, R. A., & Prance, G. T. (1974). *The design of experiments*. United Kingdom: Hafner Press.
Floridi, L., Taddeo, M., & Turilli, M. (2009). Turing's imitation game: Still an impossible challenge for all machines and some judges. *Minds and Machines, 19*(1), 145–150.
French, R. (1990). Subcognition and the limits of the Turing test. *Mind, 99*, 53–65.
French, R. (2000). The Turing test: The first 50 years. *Trends in Cognitive Sciences, 4*, 115–122.
Gonçalves, B. (2023a). The Turing test is a thought experiment. *Minds and Machines, 33*(1), 1–31.
Gonçalves, B. (2023b). *The Turing test argument*. Taylor & Francis.
Hayes, P., & Ford, K. (1995). Turing test considered harmful. In *IJCAI proceedings* (pp. 972–977).
Hodges, A. (2014). *Alan Turing: The enigma, revised edition*. London: Vintage.
Johnston, M. (1989). Dispositional theories of value. In *, 63. Proceedings of the Aristotelian society, supplementary volumes* (pp. 139–174).
Jones, C. R., & Bergen, B. K. (2024). Does GPT-4 pass the Turing test?. In *Proceedings of the 2024 conference of the north American chapter of the association for computational linguistics: Human language technologies (volume 1: long papers)* (pp. 5183–5210). https://doi.org/10.18653/v1/2024.naacl-long.290
Lewis, D. (1983). Extrinsic properties. *Philosophical Studies, 44*(2), 197–200.
Maslej, N., Fattorini, L., Perrault, R., Parli, V., Reuel, A., Brynjolfsson, E., … Clark, J. (2024). *The AI index 2024 annual report*. AI Index Steering Committee, Institute for Human-Centered AI, Stanford University.
Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences, 120*(13), Article e2215907120. https://doi.org/10.1073/pnas.2215907120
Moor, J. H. (1976). An analysis of the Turing test. *Philosophical Studies, 30*(4), 249–257.
Mumford, S. (1998). *Dispositions*. Oxford University Press.
Numerico, T. (2005). *Alan Turing e l'intelligenza delle macchine*. Franco Angeli.
Osherson, D. N., & Lasnik, H. (Eds.). (1990), *Vol. 3. An invitation to cognitive science*. Cambridge Mass: MIT Press.
Pettit, P. (1991). Realism and response-dependence. *Mind, 100*(4), 587–626.
Proudfoot, D. (2014). Turing's three senses of "emotional". *International Journal of Synthetic Emotions, 5*(2), 7–20. https://doi.org/10.4018/ijse.2014070102
Proudfoot, D. (2022). Response-dependent intelligence: Turing and the concept of mind. *Minds and Machines, 32*, 411–429. https://doi.org/10.1007/s11023-020-09534-7
Roux, S. (2011). Introduction: The emergence of the notion of thought experiments. In K. Ierodiakonou, & S. Roux (Eds.), *Thought experiments in methodological and historical contexts*. Brill. https://doi.org/10.1163/ej.9789004201767.i-233.6.
Saygin, A. P., Cicekli, I., & Akman, V. (2000). Turing test: 50 years later. *Minds and Machines, 10*(4), 463–518.
Shanahan, M. (2024). Talking about large language models. *Communications of the ACM, 67*(2), 68–79. https://doi.org/10.1145/3624724
Shieber, S. M. (1994). On Loebner's lessons. *Communications of the ACM, 37*(6), 83–85.
Shieber, S. M. (2016). Principles for designing an AI competition, or why the Turing test fails as an inducement prize. *AI Magazine, 37*(1), 91–96.
Sterrett, S. G. (2000). Turing's two tests for intelligence. *Minds and Machines, 10*(4), 541–559.
Turing, A. M. (1948). Intelligent machinery. National Physical Laboratory, UK. Reprinted in. (2004). In B. J. Copeland (Ed.), *The essential Turing* (pp. 410–432). Oxford University Press.
Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433
Turing, A. M. (1952). Can automatic calculating machines be said to think? BBC radio broadcast. (Reprinted in. (2004). In B. J. Copeland (Ed.), *the essential Turing* (pp. 494–506). Oxford University Press).
Wang, P. (2019). On defining artificial intelligence. *Journal of Artificial General Intelligence, 10*(2), 1–37. https://doi.org/10.2478/jagi-2019-0002
Weatherson, B., & Marshall, D. (2023). Intrinsic vs. extrinsic properties. In *Stanford encyclopedia of philosophy*.