Transform or combine? Tracing the irreducibility of human actions with respect to chatbot by examining definitions and evaluation tests

by *Simone D'Alessandro*[*]

Are creativity and intelligence distinct or coinciding concepts? How do they determine distinctions between humans and chatbots? Is it also possible to distinguish between incapacities? Integrating ethnomethodology and discursive analysis, the research examines the differences between human and A.I. incapacities by analysing the theoretical assumptions and international tests used by programmers to evaluate interactions: a) Turing Test; b) Winograd and Winogrande Test; c) Lovelace Test.

*Keyword*s: automatism; chatbot; creativity; incapacity; artificial intelligence; semantics.

Trasformare o combinare? Tracciare l'irriducibilità delle azioni umane rispetto ai chatbot esaminando definizioni e test di valutazione

Creatività e intelligenza sono concetti distinti o coincidenti? Come determinano le distinzioni tra esseri umani e chatbot? È possibile distinguere anche le incapacità? Integrando etnometodologia e analisi discorsiva, la ricerca esamina le differenze tra le incapacità umane e quelle delle I.A. analizzando i presupposti teorici e i test internazionali utilizzati dai programmatori per valutare le interazioni: a) Test di Turing; b) Test di Winograd e Winogrande; c) Test di Lovelace.

Parole chiave: automatismo; chatbot; creatività; incapacità; intelligenza artificiale; semantica.

Introduction: the semantic ambiguities of assumptions and research paths linking creativity and intelligence

Ethnomethodology has shown how banal or common-sense social discourses, actions, and phenomena reveal more revolutionary, creative, and counter-intuitive significance than we expect at first naive observation.

DOI: 10.5281/zenodo.17522282

[*] Università degli Studi G. D'Annunzio di Chieti-Pescara. simone.dalessandro@unich.it.

It all depends on what questions we ask, how our attention is captured, and what starting assumptions we accept, taking for granted definitions that we do not subject to further falsification (Garfinkel, 1991).

Indeed, it would always be appropriate to subject the assumptions of definitions to careful investigation in order to understand whether they favour self-deception (Zimmerman, Pollner, 1983). When researchers circumscribe definitions of words that imply complex concepts, they construct a concatenation of associations with other terms that they consider compatible with their own discursive repertoire.

Such repertoires imply cultural and ideological assumptions that constrain explicit words, generating specific constructions of meaning.

This also happens with the terms like intelligence and creativity. Definitions emerge that at first seem clear, but in the process of association with other terms, present ambiguities that scholars tend to disambiguate, selecting a reductionist path that makes the research objectifiable.

From the definitions given to intelligence, complementary relationships emerge with the term 'creativity' and its repertoires, where the ability to adapt to new situations and the capacity to transform reality in a useful and improving way are emphasised.

I will dwell specifically on the ambivalent relationships between intelligence and creativity, because from these derive distinct ways of understanding artificial intelligence, particularly that emerging from the *machine-learning* algorithms used in the large language models of chatbots (LLM). From these distinctions emerge, in turn, different ways of understanding concepts apparently opposed to the concept of intelligence: human and artificial incapability. By human inability, I mean the subjective limits of understanding of what is said or shown by a human being: in this sense, each subject has its own deficient nature.

By artificial inability I mean an objective set of limitations: 1. Inability to understand semantics; 2. Inability to contextualize conversational assumptions and implicatures; 3. Inability to interact or act unpredictably; 4. Inability to decide in the absence of starting information provided by the program; 5. Inability to boycott the automation of programming systems. Should creativity and intelligence be thought of as two distinct or coinciding aspects? The answer to this research question depends on the definitions selected by researchers. Theories always depend on social constructions and cultural ways of understanding discourse terms semantically. Anastasi and Schaefer (1971) state that creativity and intelligence are not clearly distinguishable concepts. De Caroli examines studies that support an interdependent relationship between Creativity and IQ, but also studies that

deny this correlation (De Caroli, 2016). Getzels' (1962) empirical research shows that there is a low correlation between creativity and intelligence.

Arieti (1990) and Klein (2022) confirm Getzels' thesis. Polanyi (1979), Sternberg (1988) and Sennett (2008) argue that creative processes depend on tacit knowledge, manipulation of reality and serendipity.

Hadamard (2022) emphasises the relationship between emotion and cognition. Baron-Cohen argues that autism drives specific creative and inventive processes (Baron-Cohen, 2021). Power establishes a link between creativity, schizophrenia, and bipolar disorder (Power, 2015). There are theories that distinguish intelligence from creativity and theories that start from inclusive definitions.

Researchers who start from inclusive definitions believe that creative behaviour is made up of a mix of capabilities and incapabilities. A creative subject manifests an intelligence capable of overturning the assumptions of a discourse, but incapable (or poorly capable) of being analytical.

The creative person connects distant semantic concepts (through metaphors) but is less able to connect logically close concepts.

Can we then say that, under certain conditions, incapacity can be considered a form of intelligence useful for experimenting with the creative process?

Bergson, in his work *Évolution créatrice* (Bergson, 1907; tr. it. 2012) defines automatic behaviour as 'instinct possessing its own reasons'.

Taking up this argument, social scientists have advanced the following question: what relationship does critical intelligence establish with the automatisms of routines? Is there a continuity between automatic responses and those that rebel against automatism?

There is a close correlation between the way the question is posed and the way the research path is constructed. In this essay, I will attempt to understand the relationships between combinatorial and transformative creative process by answering the following interrelated research questions: 1. Are creativity and intelligence distinct or coinciding concepts? 2. How do they determine distinctions between humans and chatbots? 3. Is it also possible to distinguish between incapacities?

I will attempt to infer clear distinctions between human and chatbot, including elements that have traditionally been discarded by researchers: ambivalences in the relationship between intelligence, creativity, inability, and automatism. In order to answer the questions posed, I will examine the tests used by programmers to distinguish human from non-human behaviour: a) the classic Turing test; b) the Winograd test and the Winogrande test; c) the Lovelace test on the creativity of artificial agents.

1. The Turing Test and the simulation of interactions: human versus artificial automatism

The Turing Test is a method for testing whether a machine, speaking through a computer interface, can be mistaken for a human being. In the test, the syntactic manifestation of verbal and para-verbal signals, emerging in the course of an interaction in the form of a dialogue (written or oral), enables the machine to simulate human thoughts, conversations and (indirectly) behaviour. The classic test was based on the relationship between three participants. Let us assume a dialogue between A and B where the two interlocutors do not see each other or cannot verify each other's identity.

We insert a third interlocutor C (human questioner). We also add that A must help C, while B (non-human) must deceive him. If C cannot distinguish A's behaviour from B's behaviour, why not attribute intelligence to B as well? (Turing, 1950). On the level of formal rules, Turing's argument is logically founded. On the other hand, if we analyse this assumption in sociological terms, the intention is evidently reductionist, as it simplifies a relationship by circumscribing it within a 'simulation of interaction', eliminating the ambivalent signals (intentional and unintentional) of non-verbal communication such as facial expressions, postures and involuntary body movements. Moreover, the test limits para-verbal expressions to the use of punctuation (in the case of textual chatbots) or to non-expressive sound intonations (in the case of vocal chatbots).

Another weakness of the Turing test, in terms of agency, is its reliance on deception as chatbots that successfully pass the Turing Test – which has been updated repeatedly to date – manage to fool humans for short periods of time, partly by evading questions (Riedl, 2014). But the crux of the debate is on how to define and, consequently, conceive the concept of comprehension.

According to a reductionist view, 'comprehension' of language simply means knowing how to use it.

According to an anti-reductionist understanding it means a) being able to grasp the nuances and latent assumptions of each sentence uttered; b) being able to contextualise the meaning of words and phrases according to the specific circumstances of a given situation and with respect to the values and cultural background of the interlocutors (Grice, 1996).

The basic issue is not related to the empirical administration of the test, but to the assumptions we make on the definitional level when we decide to give certain attributes to the terms under consideration and the resulting relationship with the alter. As the sociological literature that has dealt with

the topic shows, judging the degree of intelligence of certain behaviours depends on how we, as human beings, choose the criteria that enable the social construction of specific value judgements (Mazzotti, 2015).

If we look at the Turing test under a sociological lens, we could reverse the purpose of the test itself, stating that this method does not allow us to understand whether a machine is intelligent because it can deceive a human observer by simulating credible interactions. Instead, the test would show the extent to which humans can have automatisms that direct and influence the expectations of our conversations. Reversing Turing's perspective, we could say that the test allows us to better explain how certain habitual patterns of human beings function. Pareto (1916) had already shown the extent to which human beings act on the basis of repeated combinations.

Intelligence and creativity are related to habits and automatisms that we could also call 'action programmes' that allow us to enter into relationships without necessary self-reflection. After all, most human conversations are stereotyped and follow predictable and mechanical flows.

Think of introductory phrases, clichés, rituals, conversational turns: elements of interaction examined by the Palo Alto school, symbolic interactionism, and ethnomethodology.

The term 'automaton', which has been used since the third century BC to represent objects and artefacts that simulate the behaviour of the living, derives from a Greek term with an ambivalent meaning: *αὐτόματος*.

This term denotes an (s)object 'acting of its own will', but in common usage it connotes 'behaving like an automaton in the sense of acting mechanically and without thinking'.

Consequently, «we should not look for the automaton among machines, as we would naturally do (...) Rather, following a Bergsonian indication, its archetype should be sought among living beings» (Ronchi, 2021: 2).

Can the recurring, predictable and automatic aspects of the human (easily reproduced by an artificial intelligence) be defined as fully intelligent? If these automatisms produce something unexpected, can they be considered creative? Programs capable of generating dialogues show that there is a creative part that is recursive, i.e.: based on a few rules that can endlessly generate new productions that can have meaning and originality for the human being who observes and interprets them. In this case we can speak of repetitive and fractal meta-rules of creativity (D'Alessandro, 2023): a part that is present within an algorithm but is also possessed by the human being in terms of routines. This part is what unites the human with the artificial non-human, and we could call it combinatorial creativity (Boden, 1990). Creativity that applies rules that can create other rules to change and generate

something new in terms of a) assembling elements; b) subtraction-partition; c) reversal of data (D'Alessandro, 2025). However, while combinatorial creativity recombines prior knowledge, transformative creativity invents new categories that enable the emergence of the unexpected (Klein, 2022).

This type of transformative intelligence is related to the human capacity to interpret what is happening in a different way, reversing the perspective, or cancelling it out altogether: the person voluntarily decides to unpredictably distort an automatic path, generating paradoxes that allow established knowledge to be transcended, creating other content.

If the machine-learning algorithm generates on the basis of existing data combinations, the human being is able to 'renounce' the exploration of a topic by constructing other meanings.

The Turing test does not prove that machines can understand the non-automatic and unpredictable part of the human.

2. Winograd and Winogrande: the differences between human and artificial stupidity

Terry Winograd invented a test (still used today in his later reworkings) that demonstrates the inability of artificial intelligence, i.e., its specific inability that is very different from that of humans (Winograd, 1972).

The test is characterised by the administration of ambiguous sentences.

Winograd's test has several versions, called schemes, consisting of two sentences that differ only by one or two words, but contain an ambiguity that is resolved in opposite ways.

It is not possible to pass the test using only syntax rules. There is a need for semantic understanding and contextualisation of reality.

This is an example of Winograd's scheme: a) The city councillors refused permission to the demonstrators because they feared riots; b) The city councillors refused permission to the demonstrators because they instigated riots. People interpret the first sentence to mean that it is the city councillors who fear riots; they interpret the second sentence to mean that the instigators are the protesters. The sentences are structurally identical, but the human being, by contextualising the meaning of the two sentences and 'understanding' the roles and tasks of the city councillors and the protesters, selects the two distinct subjects. In conversations and relationships between human beings, sentences like these are frequent. When not properly understood, they give rise to misunderstandings and trigger conflicts. However, in the course of human conversation, repeated contextualisation

attempts allow misunderstandings to be resolved. Conflict may persist due to attitudes that are beyond rational comprehension (stubbornness, dislike, amusement, etc.) or that are motivated by other rational calculations (e.g., a definite intention to build controversy in order to break off relations with an interlocutor). Artificial intelligence, on the contrary, fails to contextualise and fails in an act that human beings consider simple.

There are 150 examples of Winograd schemes blocking chat bots. Because chat bots do not understand conversational implicatures[1]. Implicatures are inferences that an interlocutor makes when talking to another interlocutor, trying to understand the implicit dimensions that depend not only on the sentence, but also on the intentions and expectations of the other. They can be conventional (dependent on the shared meaning of the words used) or conversational. Conversational implicatures depend on the cooperative contribution and accepted orientations in the conversation (Grice, 1996).

In 2012, a group of researchers at New York University perfected Winograd's test by using pairs of sentences that differ by only one word (containing an object-complement pronoun that reverses the meaning of the sentence), which are followed by two questions, one for each sentence, e.g.: 1) I poured milk from the container into the jug until it was full; question: what is full, the container or the jug? 2) I poured the milk from the container into the jug until it was empty; question: what is empty, the container or the jug? With the improved Winograd, researchers confirm the semantic inability of chat boxes.

In 2019, a second group of researchers from the Allen Institute for Artificial Intelligence created 'Winogrande': a test with 44,000 sentences on a variety of topics. "While humans scored very high, the language models of the neural network (...) scored much lower"[2].

3. The Lovelace Test and exploratory creativity

The Lovelace 2.0. is a test that seeks to measure the creative capacity of a computational system, attempting to formalise the notions of originality and surprise (Bringsjord, Bello, Ferrucci, 2001).

[1] You can find the collection of such examples at:
https://cs.nyu.edu/~davise/papers/WinogradSchemas/WSCollection.html
[2] Sakaguchi K., Le Bras R., Bhagavatula C., Choi Y. (2019).

Riedl (2014) proposes an articulated test to show that a certain subset of creative acts exclusively requires human intelligence. The test is called Lovelace 2.0. Here the artificial agent is challenged on the basis of the following rules of engagement: 1. 'a' (artificial agent) must create an artefact 'o' of type 't' (output as the of processes that can be repeated and are not random hardware errors); 2. 'o' must conform to a set of constraints 'C' where ci ∈ C and is any criterion expressible also in natural language; 3. A human evaluator 'h', having chosen t and C, is satisfied that 'o' is a valid variation of t and satisfies C; 4. a human arbiter determines that the combination of t and C is not unrealistic for a human.

The constraints set make the test Google-proof and resistant to Searle's Chinese Room arguments[3]. An evaluator may impose the constraints he deems necessary to ensure that the system produces a surprising artefact or story with an original subject. Although C need not be expressed in natural language, the set of possible constraints must be equivalent to the set of all concepts that can be expressed by a human mind.

The Lovelace 2.0 Test is designed to encourage scepticism in human evaluators. This test assumes that creativity is a distinctive trait of human intelligence, but not an exclusive one.

Once again, the type of test is influenced by the starting assumptions made by the researcher. So far, human evaluators have not been surprised by machine responses. A.I.'s creativity is confirmed combinatory but does not show unpredictable and transformative behaviour.

Conclusions

As I have tried to show in my research work, the irreducibility between human intelligence and chatbot depends on the theoretical and discursive assumptions accepted by researchers investigating these fields.

I have highlighted the irreducible differences between human intelligence, creativity, automatisms and incapacities with respect to artificial expert systems, analysing theoretical assumptions and tests used by

[3] In the present paper, I have not examined Searle's philosophical experiment of the Chinese Room, presented in the article Minds, Brains and Programs, published in 1980 in the scientific journal The Behavioural and Brain Sciences, because it has not been turned into a test that can be submitted to a chat bot:
https://plato.stanford.edu/archives/win2020/entries/chinese-room/

programmers to distinguish human conversation from non-human interactions, in particular: a) The Classical Turing Test; b) The Winograd Test and The Winogrande Test; c) The Lovelace 2.0 Test.

Going back to the questions posed in the introduction, I highlight the most important factors that I understood during the research phases:

1. Are creativity and intelligence distinct or coinciding concepts? Researchers today are divided on this issue. There are theories that exclude correlations between intelligence and creativity; theories that are inclusive; theories that are ambivalent. Each starting definition generates different relationships. In some cases, these definitions have generated evaluation tests. The Turing Test assumes intelligence as the ability to emulate and dissimulate behaviour through verbal and paraverbal language, excluding the concept of comprehension and non-verbal forms of communication. The Winograd and Winogrande tests probe the semantic inability of the machine through sentences that can only be understood if they are contextualised. The Lovelace 2.0 test measures the machine's inability to 'surprise' the human evaluator. Each of these tests arises on the basis of assumptions and implications hidden in the definitions given to the terms intelligence and creativity. The way of selecting the definition determined the way of testing A.I.
2. Do Creativity and Intelligence determine irreducible distinctions between artificial human and non-human? The tests adopted by the programmers showed the presence of A.I. incapacities completely different from human ones: 1. Inability to understand semantics; 2. Inability to contextualise conversational implicatures; 3. Inability to respond or act unpredictably; 4. Inability to decide in the absence of starting information; 5. Inability to boycott the automatism of programming rules.
3. Can creativity and intelligence also imply automatism and incapacity as additional heuristic resources? Again, this all depends on how concepts are defined and what assumptions are involved in the definitions. Human beings often act automatically, demonstrating a will that would appear to be devoid of thought and equipped with habitual rules. Yet, the human's automatic mode of being is different from the artificial one in that it can be interrupted by consciousness. Moreover, the scientific community has not agreed on the concept of automatism: can what is automatic in the human and, consequently, reproducible by an artificial intelligence, be defined as intelligent? Among researchers who make distinctions

between human and artificial, the point is not technological, but cultural and ideological. The Lovelace test shows that AI is endowed with combinatorial creativity, but not transformative.

In conclusion, we can state that theoretical assumptions and tests prove that humans are more mechanical than they think they are. However, tests do not prove that machines can understand the non-mechanical part of the human. AI represents a distinct form of agency with programmable goals, data-driven adaptability, and distributed functionality. Unlike human agents, AI lacks consciousness, intentionality, and intelligence (Floridi, 2025).

The conclusions of the article can also be summarized in the following table:

Test	**Purpose**	**AI Capabilities Tested**	**Human vs AI Distinction**	**Creativity Type Involved**	**Main Limitations of AI**
Turing	Assess whether a machine can imitate human conversation	Emulation of human-like responses	Focuses on surface-level imitation, not comprehension	Combinatorial (rule-based recombination)	Lacks semantic understanding Cannot interpret non-verbal cues Relies on deception
Winograd / Winogrande	Test semantic and contextual understanding	Contextual disambiguation and implicature comprehension	Humans resolve ambiguity through world knowledge; AI fails without explicit cues	None (focus is on comprehension, not creation)	Inability to resolve ambiguity Fails to grasp conversational implicatures
Lovelace 2.0	Evaluate AI's creative capacity under constraints	Originality, surprise, and constraint satisfaction	AI shows combinatorial creativity but lacks transformative creativity	Combinatorial (rule-based generation), not transformative	Cannot surprise evaluators Lacks intentionality and unpredictability

Synoptic table: Creativity, Intelligence, and AI Evaluation Tests.

References

Arieti S. (1990). *Creatività*. Roma: Pensiero Scientifico.

Baron Cohen S. (2021). *I geni della creatività. Come l'autismo guida l'invenzione umana*. Milano: Raffaello Cortina.

Bergson H. (1907). *L'évolution créatrice*. Paris: Les Presses universitaires de France. Tr. it. (2012). *L'evoluzione creatrice*. Milano: Bur.

Bringsjord S., Bello P., Ferrucci D. (2001). Creativity, the Turing Test, and the (better) Lovelace Test. *Minds and Machines*, 11: 3-27. https://doi.org/10.1023/A:1011206622741

D'Alessandro S. (2025). *La regola che cambia le regole. Sociologia dei processi creativi e degli ecosistemi innovativi*. Milano: Mimesis.

D'Alessandro S. (2023). Creative flows: constructions of meaning between binary oppositions, paradoxes and common sense. *Italian Sociological Review*, 13(3): 371-392. https://doi.org/10.13136/isr.v13i3.668

De Caroli M.E. (2016). *Pensare, essere, fare creativamente*. Milano: FrancoAngeli.

Floridi L. (2025). AI as agency without intelligence: on artificial intelligence as a new form of artificial agency and the multiple realisability of agency thesis. *Philosophy and Technology*, 38(30): 1-30. https://doi.org/10.1007/s13347-025-00858-9

Garfinkel H. (1991). *Studies in Ethnomethodology*. Cambridge: Polity Books.

Getzels J.W. (1962). *Creativity and Intelligence*. Hoboken: John Wiley & Sons.

Grice P. (1993). *Logica e conversazione*. Bologna: il Mulino.

Hadamard J. (2022). *La psicologia dell'invenzione in campo matematico*. Milano: Raffaello Cortina.

Klein S. (2022). *Come cambiamo il mondo. Breve storia della creatività umana*. Torino: Bollati Boringhieri.

Levesque H.J., Davis E., Morgenstern L. (2012). The Winograd schema challenge. *13th International Conference on the Principles of Knowledge Representation and Reasoning*, 552-561. https://nyuscholars.nyu.edu/en/publications/the-winograd-schema-challenge-2

Mazzotti M. (2015). Per una sociologia degli algoritmi. *Rivista Italiana di Sociologia*, 3-4: 465-478. https://doi.org/10.1423/81801

Power R.A. (2015). Polygenic risk scores for schizophrenia and bipolar disorder predict creativity. *Nature Neuroscience*, 18: 953-955. https://doi.org/10.1038/nn.4040

Ronchi R. (2021). Il Bergson di Leoni. L'organo della stupidità. www.doppiozero.com/lorgano-della-stupidita

Riedl M.O. (2014). The Lovelace 2.0 test of artificial creativity and intelligence. *AAAI Symposium on Advances in Cognitive Systems*. https://doi.org/10.48550/arXiv.1410.6142

Sakaguchi K., Le Bras R., Bhagavatula C., Choi Y. (2019). *WINOGRANDE: an adversarial Winograd schema challenge at scale*. Allen Institute for Artificial Intelligence, University of Washington. https://doi.org/10.48550/arXiv.1907.10641

Searle J.R. (1990). Is the brain's mind a computer program? *Scientific American*, 262(1): 26-31. https://doi.org/10.1038/scientificamerican0190-26

Sennett R. (2008). *The Craftsman*. New Haven-London: Yale University Press.

Sternberg R.J. (1988). *The Nature of Creativity: Contemporary Psychological Perspectives*. Cambridge: Cambridge University Press.

Turing A. (1950). Computing machinery and intelligence. *Mind*, LIX(236): 433-460. https://doi.org/10.1093/mind/LIX.236.433

Zimmerman D.H., Pollner M. (1983). Il mondo quotidiano come fenomeno. In Giglioli P.P., Dal Lago A. (a cura di), *Etnometodologia*. Bologna: il Mulino.

Winograd T. (1972). Understanding natural language. *Cognitive Psychology*, 3(1): 1-191. https://doi.org/10.1016/0010-0285(72)90002-3