AISB/IACAP World Congress 2012

Birmingham, UK, 2-6 July 2012

REVISITING TURING AND HIS TEST: COMPREHENSIVENESS, QUALIA, AND THE REAL WORLD

Vincent C. Müller and Aladdin Ayesh (Editors)

Part of

ALAN TURING YEAR

2012

Foreword from the Congress Chairs

For the Turing year 2012, AISB (The Society for the Study of Artificial Intelligence and Simulation of Behaviour) and IACAP (The International Association for Computing and Philosophy) merged their annual symposia/conferences to form the AISB/IACAP World Congress. The congress took place 2–6 July 2012 at the University of Birmingham, UK.

The Congress was inspired by a desire to honour Alan Turing, and by the broad and deep significance of Turing's work to AI, the philosophical ramifications of computing, and philosophy and computing more generally. The Congress was one of the events forming the Alan Turing Year.

The Congress consisted mainly of a number of collocated Symposia on specific research areas, together with six invited Plenary Talks. All papers other than the Plenaries were given within Symposia. This format is perfect for encouraging new dialogue and collaboration both within and between research areas.

This volume forms the proceedings of one of the component symposia. We are most grateful to the organizers of the Symposium for their hard work in creating it, attracting papers, doing the necessary reviewing, defining an exciting programme for the symposium, and compiling this volume. We also thank them for their flexibility and patience concerning the complex matter of fitting all the symposia and other events into the Congress week.

John Barnden (Computer Science, University of Birmingham)
Programme Co-Chair and AISB Vice-Chair
Anthony Beavers (University of Evansville, Indiana, USA)
Programme Co-Chair and IACAP President
Manfred Kerber (Computer Science, University of Birmingham)
Local Arrangements Chair

Foreword from the Workshop Chairs

2010 marked the 60th anniversary of the publication of Turing's paper, in which he outlined his test for machine intelligence. Turing suggested that consideration of genuine machine thought should be replaced by use of a simple behaviour-based process in which a human interrogator converses blindly with a machine and another human. Although the precise nature of the test has been debated, the standard interpretation is that if, after five minutes interaction, the interrogator cannot reliably tell which respondent is the human and which the machine then the machine can be qualified as a 'thinking machine'. Through the years, this test has become synonymous as 'the benchmark' for Artificial Intelligence in popular culture.

There is both widespread dissatisfaction with the 'Turing test' and widespread need for intelligence testing that would allow to direct AI research towards general intelligent systems and to measure success. There are a host of test beds and specific benchmarks in AI, but there is no agreement on what a general test should even look like. However, this test seems exceedingly useful for the direction of research and funding. A crucial feature of the desired intelligence is to act successfully in an environment that cannot be fully predicted at design time, i.e. to produce systems that behave robustly in a complex changing environment - rather than in virtual or controlled environments. The more complex and changing the environment, however, the harder it becomes to produce tests that allow any kind of benchmarking. Intelligence testing is thus an area where philosophical analysis of the fundamental concepts can be useful for cutting edge research.

There has been recently a growing interest in simulating and testing in machines not just intelligence, but also other mental human phenomena, like qualia. The challenge is twofold: the creation of conscious artificial systems, and the understanding of what human consciousness is, and how it might arise. The appeal of the Turing Test is that it handles an abstract inner process and renders it an observable behaviour, in this way, in principle; it allows us to establish a criteria with which we can evaluate technological artefacts on the same level as humans.

New advances in cognitive sciences and consciousness studies suggest it may be useful to revisit this test, which has been done through number of symposiums and competitions. However, a consolidated effort has been attempted in 2010 and in 2011 at AISB Conventions through TCIT symposiums. However, this year's symposium forms the consolidated effort of a larger group of researchers in the field of machine intelligence to revisit, debate, and reformulate (if possible) the Turing test into a comprehensive intelligence test that may more usefully be employed to evaluate 'machine intelligence' at during the 21st century.

The Chairs
Vincent C. Müller (Anatolia College/ACT & University of Oxford) and
Aladdin Ayesh (De Montfort University)

With the Support of:
Mark Bishop (Goldsmiths, University of London),
John Barnden (University of Birmingham),
Alessio Plebe (University Messina) and
Pietro Perconti (University Messina)

The Program Committee:
Raul Arrabales (Carlos III University of Madrid),
Antonio Chella (University of Palermo),
Giuseppe Trautteur (University of Napoli Federico II),
Rafal Rzepka (Hokkaido University)
… plus the Organizers Listed Above

The website of our symposium is on http://www.pt-ai.org/turing-test

Cite as:

Müller, Vincent C. and Ayesh, Aladdin (eds.) (2012), *Revisiting Turing and his Test: Comprehensiveness, Qualia, and the Real World (AISB/IACAP Symposium)* (Hove: AISB).

Surname, Firstname (2012), 'Paper Title', in Vincent C. Müller and Aladdin Ayesh (eds.), *Revisiting Turing and his Test: Comprehensiveness, Qualia, and the Real World (AISB/IACAP Symposium)* (Hove: AISB), xx-xx.

Table of Contents

Foreword from the Congress Chairs 3

Foreword from the Workshop Chairs 4

Daniel Devatman Hromada
From Taxonomy of Turing Test-Consistent Scenarios Towards Attribution of Legal Status to Meta-modular Artificial Autonomous Agents 7

Michael Zillich
My Robot is Smarter than Your Robot: On the Need for a Total Turing Test for Robots 12

Adam Linson, Chris Dobbyn and Robin Laney
Interactive Intelligence: Behaviour-based AI, Musical HCI and the Turing Test 16

Javier Insa, Jose Hernandez-Orallo, Sergio España, David Dowe and M.Victoria Hernandez-Lloreda
The anYnt Project Intelligence Test (Demo) 20

Jose Hernandez-Orallo, Javier Insa, David Dowe and Bill Hibbard
Turing Machines and Recursive Turing Tests 28

Francesco Bianchini and Domenica Bruni
What Language for Turing Test in the Age of Qualia? 34

Paul Schweizer
Could there be a Turing Test for Qualia? 41

Antonio Chella and Riccardo Manzotti
Jazz and Machine Consciousness: Towards a New Turing Test 49

William York and Jerry Swan
Taking Turing Seriously (But Not Literally) 54

Hajo Greif
Laws of Form and the Force of Function: Variations on the Turing Test 60

From Age&Gender-based Taxonomy of Turing Test Scenarios towards Attribution of Legal Status to Meta-Modular Artificial Autonomous Agents

Daniel Devatman Hromada[123]

Abstract. The original TuringTest is modified in order to take into account the age&gender of a Judge who evaluates the machine and the age&gender of a Human with whom the machine is compared during evaluation. This yields a basic taxonomy of TuringTest-consistent scenarios which is subsequently extended by taking into account the type of intelligence being evaluated. Consistently with the Theory of Multiple Intelligences, nine basic intelligence types are proposed, and an example of a possible scenario for evaluation of emotional intelligence in early stages of development is given. It is suggested that specific intelligence types can be subsequently grouped into hierarchy at the top of which is seated an Artificial Intelligence labelled as "meta-modular". Finally, it is proposed that such a meta-modular AI should be defined as an Artificial Autonomous Agent and should be given all the rights and responsibilities according to age of human counterparts in comparison with whom an AI under question has passed the TuringTest.[123]

1 BASIC TURINGTEST TAXONOMY

Primo [1] , we present a labelling schema for a set of diverse artificial intelligence tests inspired by a procedure initially proposed by Alan Mathison Turing [2], in order to standardize attribution and evaluation of the legal status of Artificial Agents (AAs), be it robot, chat-bot or any other non-organic verbally interacting system.

For those who do not know or have forgotten what Turing Test (TT) is, we precise that according to a man rightfully labeled as a founding figure of theoretical informatics, a TT is a way how to address the question whether "Can machines think?" in a scientifically plausible yet deeply empathic way.

More concretely, Turing proposes that the performance of an AA under question shall be evaluated by a human judge (J) whose objective is to determine which among two entities -with whom J is in real-time interaction- is of human and which is of artificial nature. Traditionally, more attention was pointed upon the role of AA aiming to 'trick' J into thinking that AA is human (H). We, however, propose to partially turn the attention to roles of J & H. For it is evident that factors like J/H's age, gender or J/H's level of expertise play certain role in assessing TT's final result.

At the very core of our AA denotation schema, one finds a "TT" bigram signifying either "Turing Test", "Test Taxonomy" or whatever else one chooses them to denote. Without additional prefixes or suffixes, the presence of a TT bigram in the standardized label of an AA indicates that the candidate has already successfully passed at least one instance of the Turing Test accepted by the community. When a numeric prefix is given, it denotes the age of a judge, or age average a statistically relevant group of judges who evaluated the test. On the contrary, when a numeric postfix is given, it denotes the age of a human counterpart - or age average of human counterparts - in comparison with whom the AA has succeeded to pass the test.

A case may occur where J and|or H's gender shall significantly influence the TT-procedure (c.f. the "mother judge" example in part 3 of this article). Thus it seems to be reasonable to integrate gender information into the labeling schema.

Thus, turing tests evaluated according to the criteria proposed in paragraphs can be labeled by means of schema having the form:

$$\mathbf{G_j?jjTThhG_h?}$$

$\mathbf{G_j}$ denotes J's gender, $\mathbf{G_h}$ denotes H's gender and **jj** and **hh** tokens are substituted by age (in years) of judge or human respectively. "**?**" is a regular expression quantifier indicating that the gender information is only facultative and can be omitted for certain sets of tests.

For example, an AA which was not recognized as an artificial entity – and therefore passed the Turing Test - when compared to statistically relevant number of 18-year old human male counterparts while being evaluated by statistically significant relevant number of 21-year old female judges, shall be labelled as F21TT18M-compliant AA.

As will be stated in the last part of the article, we propose that a F21TT18M-compliant AA shall obtain certain legal rights, especially if the test concerned AA's meta-modular faculties.

After focusing attention upon J&H's age or gender, another set of variants of TT-like scenarios become possible. While in case of the standard TT, the objective is to "persuade the judge of one's human nature", in an age-oriented scenario, an agent's objective can be simply to "persuade the judge of being older than the human counterpart", while in the gender-oriented scenarios, an agent's objective can be simply to "persuade the judge that it is I and not the other player which is of sex X".

We consider it worth mentioning that the latter set of scenarios are as close as one can get to "The Imitation Game" proposed by Turing at the very beginning of his epochal article [2] and hence seems to be the closest to the origin of the TT idea.

[1] Slovak Technical University, Faculty of Electrical Engineering and Information Technology, Institute of Control and Industrial Informatics, Bratislava, Slovakia. Email: hromi@kyberia.sk.

[2] Lutin Userlab affiliated to doctoral school Cognition, Langage, Interaction of University Paris 8, France.

[3] Cognition Humaine et Artificielle laboratory (ChART) affiliated to École Pratique des Hautes Études.

2 EXTENDED TURINGTEST TAXONOMY

Secundo, we propose to crossover Turing's idea with Theory of Multiple Intelligences (TMI) articulated by Gardner [3]. According to this theory, the human intelligence is not of a general and monolithic nature, but is rather modular and task-specific. Hence, it is often the case that a human subject has an overdeveloped intelligence of one type while having the other intelligences underdeveloped. Since such is definitely the case for AA's – take for example Deep Blue's chess playing capacity and compare it to its non-existing emotional faculties – we consider the TMI to be more down-to-earth paradigm of departure for AA legal status attribution than a somewhat monolithic and language-oriented notion of intelligence held by Turing.

Inspired by TMI, we thus unfold the bTTT's labeling schema into an Extended TuringTest Taxonomy (eTTT) schema which takes into account the existence of possible intelligence types by representing them in a formalism which injects one or more infixes denoting one or more AA's intelligences between two T's of a TT-notation of of the basic schema. Thus, an extended labeling schema has the form:

$$\mathbf{G_i?jjTinThhG_h?}$$

where "in" bigram is substituted by an intelligence type infix (lowercase) from the Table 1.

Infix	Intelligence Type
em	Emotional Intelligence
li	Linguistic Intelligence
ml	Mathematico-logical Intelligence
mo	Moral Intelligence
mu	Musical Intelligence
or	Organic Intelligence
sp	Spatial Intelligence
sx	Sexual Intelligence
vi	Visual Intelligence

Table 1. Intelligence types of TMI-inspired Extended TuringTest Taxonomy and associated bigram infixes.

It is, however, more than probable that the practice and new knowledge generated especially by cognitive sciences shall indicate&recognize that more intelligence types exist. In such a case the list of basic intelligence types will have to be updated.

An example of an extended labelling schema can be an AA labelled as M42TmoT18F which succeeded to trick a statistically significant number of human 42-year-old male judges into thinking that she is at least an 18-year-old human female, and this happened under the specific constraints of "moral intelligence test " (*TmoT*) scenario. Such "*letting system answer how would it behave in a situation of ethical dilemma and why would it behave in such a way*" TmoT scenarios were already suggested by [4, 5] and myriads of others undoubtably were, are and shall be proposed for other intelligence types.

Given the diversity of forms of intelligence, the diversity of possible TT-scenarios can be astounding. We consider eTTT-like taxonomies to be of utmost importance if one does not want to get lost in this diversity during the years to come.

3 PROPOSAL FOR A F23TemT1 SCENARIO

Moral intelligence excepted, exact formulation of tests for other intelligence types is still a task to be dealt with. It is evident that in case of intelligences like sp, em or sx, a classical TT scenario based on "communication by terminal" shall not be sufficient. It may be the case that in order to evaluate such intelligence types, a human judge will have to be exposed to either physically embodied material agent in a concrete real-life situation or at least to its virtual avatar.

Let's take emotional intelligence as an example. It is a particular characteristic of emotional intelligence that it integrates information coming from and through all five corporal senses – i.e. touch with vision with audition with tastes with olphactoric inputs. We define em's 1) "active" or "production" component as an ability of an agent to express one's own internal states in a way that shall influence the state of an co-interacting counterpart; its 2) "passive" or "perception" component as an ability to parse&interpret such internal state-expressing messages coming from a co-interacting counterpart and its 3) "holistic" or "mimesis" component as an ability to adapt the active component to patterns recognised by the passive component.

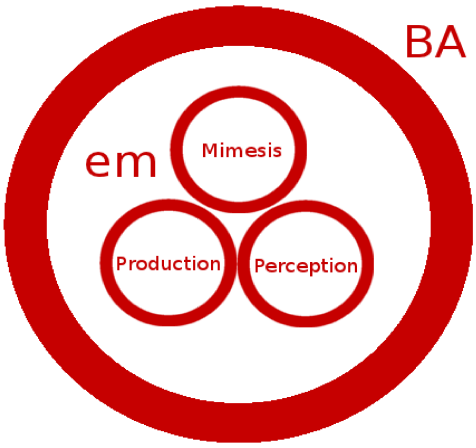


Figure 1. Triadic representation of the internal structure of emotional intelligence which is, in its turn, embedded into the substrate of the "babbling" cluster.

For those who dispose of it, visual sense seems to be the most important resource of em-relevant data and in many cases, emotional intelligence overlaps the visual one. Since we believe that the ability to recognize (passive component) and produce (active component) facial expressions is an integral part of em, we state that a TT-consistent test evaluating these faculties is closer to an ideal "Turing-like emotional intelligence test" (TemT), than a test which does not evaluate such faculties. But, of course, AA's facial-expression production capabilities cannot be evaluated if an AA that does not "have a face" at all!

Therefore an AA under question should "obtain a face" - purely virtual avatar should be sufficient in the first versions of TemT, but an embodiment in a material robotic face shall be considered as necessary in later stages, when one would aim to integrate agent's emotional intelligence with especially its somato-sexual counterpart (*sx*).

With regards to visual aspects of *em*, we add that what was said about facial expressions applies, *mutatis mutandis*, to bodily gestures & movements .

em's second most important input/output channel is the auditive channel. With regards to these considerations, emotional intelligence should be placed in between the musical and linguisticl intelligence. For in majority of cases it is not the syntactic structure or the semantic content of the linguistic utterance which has the most immediate effect upon the hearer, it is the prosodic layer – the rhythm, the tempo, the tone&trembling of voice - which carries a "direct" emotional message.

On the contrary, possibly emotionally charged reactions to text strings displayed on terminal in classic TT-scenario can be considered only as "indirect" since from strictly pragmatic point of view, they are nothing else than graphematic consequences of semantic interpretation of the displayed utterance.

Combination of the visual and auditive modalities could be possibly sufficient for tests evaluating whether emotional intelligence of an AA can be considered as equivalent to that of a human child. One can imagine a test during which a 23-year old Judge observes & interacts – by means of microphones, speakers, web-cams and touch screens - with two "babies" and tries to identify which one is human and which one is not.

While the human baby is as human as it can be, with its fingers touching those tablet regions where Judge's smiling face is displayed, it may be the case that an artificial "baby" does not have any material essence at all. But does not mean that Judge will perceive it to be less natural than the human baby: with a robust smile detector [6], a good underlying wire-frame facial model and a algorithm for "pleasing cry" synthesis it may well be considered more "human" a baby than a human baby and thus pass a F23TemT1 test.

This example of a F23TemT1 which can be theoretically realised using current technologies emphasises the importance of J & H roles in execution of TT-like scenarios. For it seems to be quite evident that a human baby suffering an autistic disorder complicating his perception of smiles shall perform less "naturally" than a normal baby and most probably also less naturally than an AA furnished with a well-tuned smile detector.

It seems to be also very probable that the performance of a 23-year young mother judge in such a F23TemT1 will differ from the performance of a young man of the same age who never saw such a small baby before. We consider this *mother judge gedankenexperiment* to be a sufficiently persuasive argument for integration of gender-specific information into bTTT and eTTT labeling schemas – an argument which even the most ardent feminist shall defeat only with greatest difficulties.

Finally, it is important to mention that a scenario hereby proposed is only schematic and incomplete. It does not, for example, involve any kind of communication by means of touches, tastes and smells which is crucial in true human interaction and crucial^2 (crucial squared) in a mother-baby interaction. However, even in this highly simplified way, a F23TemT1 scenario extends the applicability of Turing's initial proposal into auditive and visual domain. This F23TemT1 can also indicate how a F23TemT2 scenario could possibly look like: F23TemT2 = F23TemT1 + little bit of Freud + lot of Piaget.

These are, according to our opinion, constructive steps towards deployment and|or emergence of an *autonomous* AA (AAA).

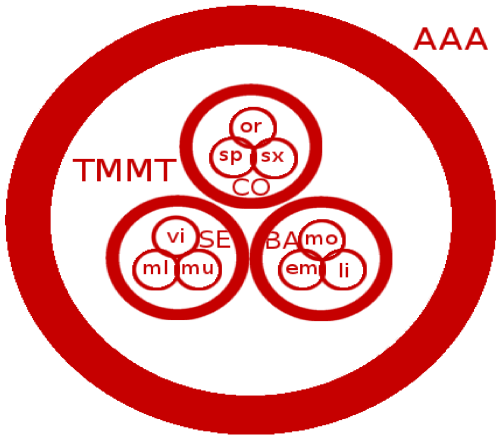
4 eTTT HIERARCHIES

The advantage of a TMI-based taxonomy of Turing Test is that every partial scenario which evaluates a specific intelligence type shall - once operationalized, standardized and agreed upon by AI & cognitive science communities – pose a set of specific and exact constraints. These constraints can be transformed into problems, and these problems can be subsequently dealt with, one by one. Thus, step after step, component after component, brick after brick and a test after test, a certain AA could be possibly raised towards the state where it shall integrate all intelligences and become a meta-modular and hence autonomous AI.

A meta-modular AI as an AI capable of passing a meta-modular TT. A meta-modular TT is a TT integrating all Turing Test scenarios associated with a given age group. Failure to pass any specific TT-scenario or failure to pass any combination of specific TT-scenarios is sufficient a reason for refusal of the label "meta-modular" to an AAA under question. This applies also for new tests as well – if a TMMT is unable to pass a TT-scenario which was just canonized by the scientific community, it can not be labeled as "meta-modular" until the moment when it shall be able to pass the test.

A meta-modular TT is at the very top of the hierarchy of TT-consistent scenarios. MM can be defined as "group of all groups", it integrates [7] all intelligence types. Underneath this layer0 is located a layer of individual intelligence groups. This is so because the formalism hereby proposed does not forbid to group intelligence types into clusters of these types according to mutual overlap of these types. And because it is neither forbidden to divide world in triads and because it can be sometimes even convenient, we hereby define these three basic intelligence clusters|groups:

BA - or a babbling cluster|group = mo * li * em
SE - or a sensual cluster|group = mu * vi * ml
CO - or a carnal cluster|group = or * sp * sx



represented also in Figure 2 and Table 2.

Figure 2. Triadic fractal chintamani representation of a possible embeddement of ETT intelligence types into intelligence clusters which are, in their turn, embedded into substrate of a meta-modular TT.

ETT infix	Intelligence group	Subordinated intelligence types
CO	Corporal group	Organic; Spatial; Somato-sexual
BA	Babbling group	Moral; Emotional; Linguistic
SE	Sensual group	Mathematico-logical; Musical; Visual

Table 2. Clustering of basic intelligence types into basic intelligence clusters

Voilà three representation of one type of possible clusterings of proposed intelligence types. Purely aesthetic reasons excepted, other reasons why we propose this and not other way of clustering TMI-inspired intelligences are as follows:

The "carnal group" consists of intelligences associated to corporal aspects of one's existence. In case of a human being, intelligences of this group are much more "innate" and in lesser extent "acquired" from environment than is the case in other two clusters. We precise: organic intelligence or (also called "natural intelligence in Gardner's model or circuits 1&2 in Leary-Wilson's model [8]) is tightly related to organism's ability to survive. Find food which you can digest, escape a predator, make a nest – all this activities are related to organic intelligence. It is also especially organic intelligence which deals with nociceptive, olphactoric and gustative inputs. Somato-sexual intelligence sx assures reproduction and by activating the *qualia of pleasure* assures the survival of species and life in general. Spatial intelligence sp involves not only agent's movement within the 3D quasieuclidean space of the shared everyday reality, but also takes into account the experience of one's *body as a space* which is related especially to haptic and proprioceptive inputs.

To summarize: intelligence types of CO group require that an agent has a material body. The performances of this body can be tested by means of many possible TT-scenarios be it danse (sp), tantryogic rite (sx) or "go&survive in a forest for a day, boy..." (or).

The BA cluster envelops those intelligence types which develop especially in early childhood of a human baby. What's more, all three intelligences of the cluster stem from the notion that there are other beings in the shared world, and that one should communicate with them (li), shall not hurt them (mo) and, if possible, try to understand them (em).

We consider it worth mentioning that during the ontogeny of an individual being, both linguistic and moral intelligence are possibly subjects to somewhat similar inductive procedures: be it "moral"[4] or "grammar" induction during which a set of general rules is being infered from positive example set encountered in the being's environment.

Finally, the SE cluster unifies those intelligence types which develop relatively lately during one's ontogeny. Verily, it is especially in domains of visual (vi), musical (mu) and mathematico-logical (ml) intelligences that one encounters *genial* performances. For further inspiration coming from this domain of Godels, Eschers and Bachs, it is worth (re)reading [9].

This kind of hierarchy we present hereby in order to prepare as solid basements as we are capable of for possible attribution of civic rights to future TAAATs.

5 ATTRIBUTION OF CIVIC RIGHTS TO AAA

Imagine an F18TmmT18F artificial agent, i.e. an artificial agent which has succeeded to persuade 18 year old women to consider her as being EITHER older OR more feminine OR more human in all possible TT-scenarios as well as in their combinations. Could such an entity be granted rights equal to rights of a defeated age group in spite of fact that the entity under question is of artificial, non-organic origin?

From strictly legal point of view, the answer could theoretically be "yes" within the framework of Rome-originated legal systems since non-organic entities like "corporations" or other "legal personnae" already dispose of certain rights within such legal frameworks.

Thus, the path towards attribution of legal rights (and responsabilities) to AAs we are programming is relatively straightforward:

1) Define the set of TTs which evaluate diverse faculties of human mind
2) Canonize them into ISO-like standard which attributes a certain label (AAA) to an agent which passes them
3) Persuade the public that it is in their own interests that entities which have obtained an AAA label shall be given at least certain civil rights

Subsequently, it shall only be necessary to pass a law or constitutional amendment stating that:

Rights & Responsabilities of AAAs are identic to their human counterparts.

.

..and the rights shall be given.

We precise that within our labeling schemas, an AA obtains the rank AAA if and only if it passes TMMT in a test where age of judge and human counterpart is identic. If it happens, we propose to consider her|him as equal among equals.

For example, in possible future liberal societies whose civic or traditional roots allow them to do so, such a 21TmmT21|TAAAT21 could possibly have a right to be granted access to adult-only discussion forum or social network, while a 17TmmT17|TAAAT17-compliant system shall not dispose of such a right. However, possibly even a 15TmmT15|TAAAT15 could, in such societies, possibly have a right to dispose of a bank account in order to execute financial transactions according to its embedded value system.

While we do not feel at all apt to answer the questions "whether such 21TmmT21|TAAAT21 or shall possibly dispose of legal rights equivalent to those of his human adult counterparts?" we nonetheless think that the value of *TT-notation* resides in the fact that it facilitates the process of posing such questions.

[4] Redaction of our "moral induction" proposal is in process .

And in the world where AI is gaining momentum, we feel that such questions should be posed.

6 SUMMARY

In order to facilitate & synchronize exchange between AI-engineers, we propose basic basic TuringTestTaxonomy (bTTT) and extended TuringTestTaxonomy (eTTT) labeling schemas.

BTTT labels can be matched by a PERL-compatible regular expression:

/([FM])?(\d+)TT(\d+)([FM])?/

where first (facultative) group matched by parenthesis group-matching operator denotes the gender of the judge(s) who performed Turing's Test, whereby second group informs of their age, the third group matches the age of human counterpart(s) and the last one (facultative) of their gender.

ETTT labels can be matched by a PERL-compatible regular expression:

/([FM])?(\d+)T([a-Z]{2})T(\d+)([FM])?/

where first (**?**=facultative) group matched by parenthesis group-matching operator denotes the gender of the judge(s) who performed Turing's Test, the second group informs of their age, the third group matches the age of human counterpart(s) and the last one (**?**=facultative) of their gender. If the third group -i.e. a bigram infix located between two Ts - is in the lower case, it denotes the intelligence type which was tested while the bigram infix in upper case indicates that the test involved cluster of intelligence types.

Nine intelligence types proposed for eTTT are enumerated in Table 1. The fact that they are initially clustered into three clusters (BA, CO, SE) within the scope of this article does not exclude other -potentially sounder- clusterings to be proposed.

The "meta-modular" MM bigram indicates that the agent under question passed all known tests as well as all their combinations. If ever a new test is accepted by a scientific community and an artificial agent previously labeled as TMMT fails to pass such a test, it shall not be considered as "meta-modular" until the moment when (s)he shall (re)?organize (him|her)self in such a way that (s)he shall pass the new test.

An AAA label can possibly be attributed to an artificial agent who had achieved such a level of autonomy [10] that it passed a TMMT in condition where age (and possibly gender) of judges is identic to age (respectively gender) of human counterparts. Since J's & H's age ares identic, for such an AAA the label can be abbreviated so that for example 21TMMT21 becomes simply TAAAT21 or even AAA21.

Formally there is no legal obstacle in attributing certain civil rights to AAAs since the most common legal frameworks already give certain rights to non-organic legal personnae [11]. The path to such attribution is relatively straightforward: define what exactly AAA means by canonizing different tests (and their combinations) in a form of a supranational standard (e.g. ISO). Afterwards, it suffices to integrate the statement like "Rights & Responsabilities of AAAs are identic to their human counterparts" into local legal codex.

Finally, to prevent possible confusion, we consider it important to state that while the question whether the labeling schemas like bTTT or eTTT can define what *rights shall be given* to a non-organic agent; under any circumstance whatsoever it is not acceptable to concieve nor apply such a legal framework which would exploit the schema hereby proposed in order to state what rights should be *taken* from an organic agent. For any such tentative would be contrary to positive and constructive intention behind this proposal and shall therefore nullify the validity of the contract between men&machines hereby proposed [12] .

ACKNOWLEDGMENTS

This article would not be written without intellectual support from doc. Sekaj and other members of URPI FEI STU team, as well as without kind guidance of prof. Tijus who has helped me to glue it all together. Many thanks go also members of MobenFact for being as inspirative as only they can be and to Adil ElGhali for initiation into semantic vector spaces.

REFERENCES

[1] K. Čapek. R.U.R. - *Rossumovi Univerzální Roboti*. Aventinum, Prague, Československá Republika. (1920).

[2] A. M. Turing. Computing Machinery and Intelligence. *Mind LIX*, 236: 433–460 (1950).

[3] H. Gardner. *Frames of mind: The Theory of Multiple Intelligences*. Basic Books. New York, USA (1983).

[4] C. Allen, G. Varner and J. Zinser. Prolegomena to Any Future Artificial Moral Agent. *J. Expt . Theor . Artif .Intell.* 12: 251-261 (2000).

[5] D. D. Hromada. The Central Problem of Roboethics: from Definition towards Solution. Procs. Of 25th annual conference of International Association for Computing and Philosophy (IACAP), Aarhus, Denmark. (2011).

[6] D. D. Hromada, C. Tijus, S. Poitrenaud and Nadel J. Zygomatic Smile Detection: The Semi-Supervised Haar Training of a Fast and Frugal System. Procs. of IEEE RIVF2012 conference. Hanoi, Vietnam. (2010).

[7] G. Tononi. An Information Integration Theory of Consciousness. *BMC Neuroscience* . 5**:**42. (2004).

[8] R. A. Wilson. *Quantum Psychology.* New Falcon. (1979).

[9] D. Hofstadter. *Godel, Escher, Bach – an Eternal Golden Braid*. Basic Books. (1979).

[10] I. Kant. *Grundlegung zur Metaphysik der Sitten*. Deutschland. (1785).

[11] I. Asimov. *Forward the Foundation*. Spectra. USA. (1993).

[12] D. D. Hromada. From Age&Gender-based Taxonomy of Turing Test Scenarios towards Attribution of Legal Status to Meta-Modular Artificial Autonomous Agents. Accepted for CQRW symposium of AISB/IACAP World Congress. Birmingham, UK. (2012).

Keyphrases. *Age. Gender. Intelligence Types. Turing Test Taxonomy. Theory of Multiple Intelligences. Meta-modular intelligence. Autonomous Artificial Agent. Attribution of civil rights to informatic systems. Chintamani fractal model of intelligence type&component clustering. BTTT and ETTT techspiecies annotation schemas* [5]

[5] Note the author to editors: Please do not delete this ***keyphrase*** *list* from the published version.

My Robot is Smarter than Your Robot - On the Need for a Total Turing Test for Robots

Michael Zillich[1]

Abstract. In this position paper we argue for the need of a Turing-like test for robots. While many robotic demonstrators show impressive, but often very restricted abilities, it is very difficult to assess how intelligent such a robot can be considered to be. We thus propose a test, comprised of a (simulated) environment, a robot, a human tele-operator and a human interrogator, that allows to assess whether a robot behaves as intelligently as a human tele-operator (using the same sensory input as the robot) with respect to a given task.

1 INTRODUCTION

The Turing Test [35] considered the equivalent of a brain in a vat, namely an AI communicating with a human interrogator solely via written dialogue. Though this did not preclude the AI from having acquired the knowledge that it is supposed to display via other means, for example extended multi-sensory interactions within a complex dynamic environment, it did narrow down what is considered as relevant for the display of intelligence.

Intelligence however encompasses more than language. Intelligence, in all its flavours, developed to provide a competitive advantage in coping with a world full of complex challenges, such as moving about, manipulating things (though not necessarily with hands), hiding, hunting, building shelter, caring for offspring, building social contacts, etc. In short, intelligence needs a whole world to be useful in, which prompted Harnad to propose the Total Turing Test [19], requiring responses to all senses not just formatted linguistic input. Note that we do not make an argument here about the best approach to explain the emergence of intelligence (though we consider it likely that a comprehensive embodied perspective will help), but only about how to measure intelligence without limiting it to only a certain aspect.

The importance of considering all aspects of intelligence is also fully acknowledged in robotics, where agents situated in the real world are faced with a variety of tasks, such as navigation and map building, object retrieval, or human robot interaction, which require various aspects of intelligence in order to be successfully carried out in spite of all the challenges of complex and dynamic scenes. So robotics can serve as a testbed for many aspects of intelligence. In fact it is the more basic of the above aspects of intelligence that still pose major difficulties. This is not to say that there was no progress over the years. In fact there are many impressive robot demonstrators now displaying individual skills in specific environments, such as bipedal walking in the Honda Asimo [6] or quadruped walking in the Boston Dynamics BigDog[32], learning to grasp [25, 33], navigation in the Google Driverless Car or even preparing pancakes [11]. For many of these demonstrators however it is easy to see where the limitations lie and typically the designers are quick to admit that this sensor placement or that choice of objects was a necessary compromise in order to concentrate on the actually interesting research questions at hand.

This makes it difficult however to quantitatively compare the performance of robots. Which robot is smarter, the pancake-flipping robot in [11][2] , the beer-fetching PR2[3] or the pool-playing PR2[4]? We will never know.

A lot of work goes into these demonstrators, to do several runs at conferences or fairs and shoot videos, before they are shelved or dismantled again, but it is often not clear what was really learned in the end; which is a shame, because certainly some challenges were met with interesting solutions. But the limits of these solutions were not explored within the specific experimental setup of the demo.

So what we argue for is a standardised, repeatable test for complete robotic systems. This should test robustness in basic "survival" skills, such as not falling off stairs, running into mirrors or getting caught in cables, as well as advanced tasks, such as object search, learning how to grasp or human-robot interaction including natural language understanding.

2 RELATED WORK

2.1 Robot Competitions

Tests are of course not new in the robotics community. There are many regular robot challenges which have been argued to serve as benchmarks [12], such as RoboCup [24] with its different challenges (Soccer, Rescue, @Home), the AAAI Mobile Robot Competitions [1], or challenges with an educational background like the US FIRST Robotics Competitions [8] or EUROBOT [3]. Furthermore there are specific targeted events such as the DARPA Grand Challenges 2004 and 2005 and DARPA Urban Challenge 2007 [2]. While these events present the state of the art and highlight particularly strong teams, they only offer a snapshot at a particular point in time. And although these events typically provide a strict rule book, with clear requirements and descriptions of the scenarios, the experiments are not repeatable and the test arena will be dismantled after the event (with the exception of simulations of course). So while offering the ultimate real-world test in a challenging and competitive setting, and thus providing very important impulses for robotics research, these tests are not suitable because a) they are not repeatable, b) rules keep changing to increase difficulty and maintain a challenging competition and c) the outcomes depend a lot on factors related

[1] Vienna University of Technology, Austria, email: zillich@acin.tuwien.ac.at

[2] www.youtube.com/watch?v=4usoE981e7I
[3] www.willowgarage.com/blog/2010/07/06/beer-me-robot
[4] www.willowgarage.com/blog/2010/06/15/pr2-plays-pool

to the team (team size and funding, quality of team leadership) rather than the methods employed within the robot.

2.2 Robotic Benchmarks

The robotics community realised the need for repeatable quantitative benchmarks [15, 21, 26, 27], leading to a series of workshops, such as the Performance Metrics for Intelligent Systems (PerMIS) or Benchmarks in Robotics Research or the Good Experimental Methodology in Robotics series, and initiatives such as the EURON Benchmarking Activities [4] or the NIST Urban Search And Rescue (USAR) testbed [7].

Focusing on one enabling capability at a time, some benchmarks concentrate on path planning [10], obstacle avoidance [23], navigation and mapping [9, 13], visual servoing [14], grasping [18, 22] or social interaction [34, 20]. Taking into account whole robotic systems [16] propose benchmarking biologically inspired robots based on pursuit/evasion behaviour. Also [29] test complete cognitive systems in a task requiring to find feeders in a maze and compete with other robots.

2.3 Robot Simulators

Robotics has realised the importance of simulation environments early on, and a variety of simulators exist. One example is Player/Stage [17], a robot middleware framework and 2D simulation environment intended mostly for navigation tasks and its extension to a full 3D environment with Gazebo [5], which uses a 3D physics engine to simulate realistic 3D interactions such as grasping and has recently been chosen as the simulation test bed for the DARPA Robotics Challenge for disaster robots. [28] is another full 3D simulator, used e.g. for simulation of robotic soccer players. Some simulators such as [30] and [36] are specialised to precise simulation of robotic grasping. These simulators are valuable tools for debugging specific methods, but their potential as a common testbed to evaluate complete robotic systems in a set of standardised tasks has not been fully explored yet.

In summary, we have on the one hand repeatable, quantitative benchmarks mostly tailored to sub-problems (such as navigation or grasping) and on the other hand competitions testing full systems at singular events, where both of these make use of a mixture of simulations and data gathered in the real world.

3 THE TOTAL TURING TEST FOR ROBOTS

What has not fully emerged yet however is a comprehensive test suite for complete robotic systems, maintaining a clearly specified test environment plus supporting infrastructure for an extended period of time, allowing performance evaluation and comparison of different solutions and measuring their evolution over time is What this test suite should assess is the overall fitness of a robotic system to cope with the real world and behave intelligently in the face of unforeseen events, incomplete information etc. Moreover the test should ideally convey its results in an easily accessible form also to an audience beyond the robotics research community, allowing other disciplines such as Cognitive Science and Philosophy as well as the general public to assess progress of the field, beyond eye-catching but often shallow and misleading demos,

Harnads [19] Total Turing Test provides a fitting paradigm, requiring that *"The candidate [the robot] must be able to do, in the real world of objects and people, everything that real people can do, in a way that is indistinguishable (to a person) from the way real people do it."*

"Everything" will of course have to be broken down into concrete tasks with increasing levels of difficulty. And the embodiment of the robot will place constraints on the things it can do in the real world, which has to be taken into account accordingly.

3.1 The Test

The test would consist of a given scene and a set of tasks to be performed by either an autonomous robot or a human tele-operating a robot (based on precisely the same sensor data the robot has available, such as perhaps only a laser ranger and bumpers). A human interrogator would assign tasks to the robot, and also place various obstacles that interfere with successful completion. If the human interrogator can not distinguish the performance of the autonomous robot from the performance of the tele-operated robot, the autonomous robot can be said to be intelligent, with respect to the given task.

Concretely the test would have to consist of a *standardised environment* with a *defined set of tasks*, as is e.g. common in the RoboCup@Home challenges (fetch an item, follow a user). The test suite would provide a API, e.g. based on the increasingly popular Robot Operating System (ROS) [31], allowing each robot to be connected to it, with moderate effort. Various obstacles and events could be made to interfere with execution of these tasks, such as cables lying on the floor, closed glass doors, stubborn humans blocking the way. Different challenges will pose different problems for different robots. E.g. for the popular omnidirectional drives of holonomic bases such as the Willow Garage PR2 cables on the floor represent insurmountable obstacles, while other robots will have difficulties navigating in tight environments.

3.2 Simulation

A basic building block for such a test suite is an extension of available simulation systems to allow fully realistic simulation of all aspects of robotic behaviour.

The simulation environment would have to provide *photo-realistic rendering with accurate noise models* (such as lens flares or poor dynamic range as found in typical CCD cameras) beyond the visually pleasing but much to "clean" rendering of available simulators. Also the *physics simulation* will have to be very realistic, which means that the simulation might not be able to run in real time. Real time however is not necessarily a requirement for a simulation as long as computation times of employed methods are scaled in accordance. Furthermore the simulation would need to also contain humans, instructing the robot in natural language, handing over items or posing as dynamic obstacles for navigation.

Figure 1 shows a comparison of a robot simulated (and in this case tele-operated) in a state of the art simulator (gazebo) with the corresponding real robot carrying out the same task autonomously as part of a competition [37]. While the simulation could in this case provide reasonably realistic physics simulation (leading to objects slipping out of the hand if not properly grasped) and simulation of sensors (to generate e.g. problems for stereo reconstruction in low-texture areas) more detailed simulations will be needed to capture more aspects of the real world.

(a) (b)

Figure 1. Comparison of (tele-operated) simulation and (autonomous) real robot in a fetch and carry task.

Task \ Stage	Simulated Environment			Real Environment		
	1: Perfect	2: Mild difficulties	3: Severe difficulties	4: Perfect	5: Mild difficulties	6: Severe difficulties
Random navigation	100 / 100	73 / 56	15 / 4	98 / 92	69 / 44	11 / 4
Navigation to a goal	97 / 95	53 / 42	10 / 5	96 / 88	43 / 40	5 / 2
Follow human	98 / 99	87 / 58	23 / 8	94 / 91	76 / 55	12 / 7
Guide human	-	-	-	-	-	-
Open door	-	-	-	-	-	-
Find object	87 / 76	44 / 24	24 / 16	74 / 63	37 / 21	16 / 4
Retrieve object	-	-	-	-	-	-
...	...	...	...	...	...	...

Figure 2. Example score for a fictional robot equipped with a laser ranger and camera, but no arm and language capabilities. Figures are scores on the Pass/fail test and Intelligence test respectively.

3.3 Task and Stages

The test would be set up in different *tasks* and *stages*. Note that we should not require a robot to do everything that real people can do (as originally formulated by Harnad). Robots are after all designed for certain tasks, requiring only a specific set of abilities (capable of language understanding, equipped with a gripper, ability to traverse outdoor terrain, etc.). And we are interested in their capabilities related to these tasks. The constraints of a given robot configuration (such as the ability to understand language) then apply to the robot as well as the human tele-operator.

Stages would be set up with increasing difficulties, such that a robot can be said to be stage-1 safe for the fetch and carry task (all clean, static environment) but failing stage 2 in 20% of cases (e.g. unforeseen obstacles, changing lighting). The final stages would be a real world test in a mock-up constructed to follow the simulated world. While the simulation would be a piece of software available for download, the real world test would be held as an annual competition much like RoboCup@Home, with rules and stages of difficulty according to the simulation. Note that unlike in RoboCup@Home these would remain fixed, rather than change with each year.

3.4 Evaluation

The test would then have two levels of evaluation.

Pass/fail test This evaluation would simply measure the percentage of runs where the robot successfully performs a task (at a given stage). This would be an automated assessment and allows developers to continuously monitor progress of their system.

Intelligence test This would be the actual Total Turing Test with humans interrogators assessing whether a task was performed (successfully or not) by a robot or human tele-operator. The score would be related to the percentage of wrong attributions (i.e. robot and tele-operator were indistinguishable). Test runs with human tele-operators would be recorded once and stored for later comparison of provided robot runs. The requirement of collecting statistics from several interrogators means that this test is more elaborate and would be performed in longer intervals such as during annual competitions. This evaluation then allows to assess the intelligence of a robot (with respect to a given task) in coping with the various difficulties posed by a real environment.

The setup of tasks and stages allows to map the abilities of a given robot. Figure 2 shows the scores of a fictional robot. The robot is equipped with a laser ranger and camera and can thus perform the navigation tasks as well as following a human, but lacks an arm for carrying objects or opening doors as well as communication capabilities required for the human guidance task,

As can be seen the robot can be considered stage-1 intelligent with respect to the random navigation task (driving around randomly without colliding or getting stuck), i.e. it is indistinguishable from a human tele-operator driving randomly, in the perfect simulated environment. It also achieves perfect success rates in this simple setting. Performance in the real world for perfect conditions (stage 4) is slightly worse (the simulation could not capture all the eventualities of the real world, such as wheel friction). Performance for added difficulties (such as small obstacles on the floor) decreases, especially in the real word condition. Performance drops in particular with respect to the tele-operator and so it becomes quickly clear to the interrogators which is the robot and which the tele-operator, i.e. the robot makes increasingly "stupid mistakes" such as getting stuck when there is an obvious escape. Accordingly the intelligence score drops quickly. The robot can also be said to be fairly stage-1 and stage-4 intelligent with respect to navigation and human following, and slightly less intelligent with respect to finding objects.

In this respect modern vacuum cleaning robots (the more advanced versions including navigation mapping capabilities) can be considered intelligent with respect to the cleaning task, as their performance there will generally match that of a human tele-operating such a robot. For more advanced tasks including object recognition, grasping or dialogue the intelligence of most robots will quickly degrade to 0 for any stages beyond 1.

4 CONCLUSION

We proposed a test paradigm for intelligent robotic systems, inspired by Harnads Total Turing Test, that goes beyond current benchmarks and robot competitions. This test would provide a pragmatic definition of intelligence for robots, as the capability to perform as good as a tele-operating human for a given task. Moreover, test scores would be a good indicator whether a robot is ready for the real world, i.e. is endowed with enough intelligence to overcome unforeseen obstacles and avoid getting trapped in "stupid" situations.

There are however several technical and organisational challenges to be met. Running realistic experiments will require simulators of considerably improved fidelity. But these technologies are becoming increasingly available thanks in part to the developments in the gaming industry. Allowing researchers to simply plug in their systems will require a careful design of interfaces to ensure that all capabilities are adequately covered. The biggest challenge might actually be the definition of environments, tasks and stages. This will have to be a community effort and draw on the experiences of previous benchmarking efforts.

ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement No. 215181, CogX and from the Austrian Science Fund (FWF) under project TRP 139-N23 InSitu.

REFERENCES

[1] AAAI Mobile Robot Competition, http://www.aaai.org/Conferences/AAAI/2007/aaai07robot.php.
[2] DARPA Grand Challenge, http://archive.darpa.mil/grandchallenge.
[3] Eurobot, http://www.eurobot.org.
[4] EURON Benchmarking Initiative, www.robot.uji.es/EURON/en/index.html.
[5] Gazebo 3D multi-robot simulator http://gazebosim.org.
[6] Honda ASIMO, http://world.honda.com/ASIMO.
[7] NIST Urban Search And Rescue (USAR), http://www.nist.gov/el/isd/testarenas.cfm.
[8] US First Robotics Competition, www.usfirst.org.
[9] Benjamin Balaguer, Stefano Carpin, and Stephen Balakirsky, 'Towards Quantitative Comparisons of Robot Algorithms: Experiences with SLAM in Simulation and Real World Systems', in *IROS Workshop on Benchmarks in Robotics Research*, (2007).
[10] J Baltes, 'A benchmark suite for mobile robots', in *Intelligent Robots and Systems 2000IROS 2000 Proceedings 2000 IEEERSJ International Conference on*, volume 2, pp. 1101–1106. IEEE, IEEE, (2000).
[11] Michael Beetz, Ulrich Klank, Ingo Kresse, Lorenz Maldonado, Alexis Mösenlechner, Dejan Pangercic, Thomas Rühr, and Moritz Tenorth, 'Robotic Roommates Making Pancakes', in *11th IEEE-RAS International Conference on Humanoid Robots*, (2011).
[12] S Behnke, 'Robot competitions - Ideal benchmarks for robotics research', in *Proc of IROS2006 Workshop on Benchmarks in Robotics Research*. Citeseer, (2006).
[13] Simone Ceriani, Giulio Fontana, Alessandro Giusti, Daniele Marzorati, Matteo Matteucci, Davide Migliore, Davide Rizzi, Domenico G Sorrenti, and Pierluigi Taddei, 'Rawseeds ground truth collection systems for indoor self-localization and mapping', *Autonomous Robots*, **27**(4), 353–371, (2009).
[14] Enric Cervera, 'Cross-Platform Software for Benchmarks on Visual Servoing', in *IROS Workshop ong Benchmarks in Robotics Research*, (2006).
[15] R. Dillmann, 'Benchmarks for Robotics Research', Technical report, EURON, (2004).
[16] Malachy Eaton, J J Collins, and Lucia Sheehan, 'Toward a benchmarking framework for research into bio-inspired hardware-software artefacts', *Artificial Life and Robotics*, **5**(1), 40–45, (2001).
[17] Brian P Gerkey, Richard T Vaughan, and Andrew Howard, 'The Player / Stage Project : Tools for Multi-Robot and Distributed Sensor Systems', in *International Conference on Advanced Robotics (ICAR)*, pp. 317–323, (2003).
[18] Gerhard Grunwald, Christoph Borst, and J. Marius Zöllner, 'Benchmarking dexterous dual-arm/hand robotic manipulation', in *IROS Workhop onPerformance Evaluation and Benchmarking for Intelligent Robots and Systems*, (2008).
[19] S Harnad, 'Other Bodies, Other Minds: A Machine Incarnation of an Old Philosophical Problem', *Minds and Machines*, **1**, 43–54, (1991).
[20] Zachary Henkel, Robin Murphy, Vasant Srinivasan, and Cindy Bethel, 'A Proxemic-Based HRI Testbed', in *Proceedings of the Performance Metrics for Intelligent Systems Workshop (PerMIS)*, (2012).
[21] I Iossifidis, G Lawitzky, S Knoop, and R Zöllner, 'Towards Benchmarking of Domestic Robotic Assistants', in *Advances in Human Robot Interaction*, eds., Erwin Prassler, Gisbert Lawitzky, Andreas Stopp, Gerhard Grunwald, Martin Hägele, Rüdiger Dillmann, and Ioannis Iossifidis, volume 14/2004 of *Springer Tracts in Advanced Robotics {STAR}*, chapter 7, 97–135, Springer Press, (2005).
[22] R. Jäkel, R., Schmidt-Rohr, S. R., Lösch, M., & Dillmann, 'Hierarchical structuring of manipulation benchmarks in service robotics', in *IROS Workshop on Performance Evaluation and Benchmarking for Intelligent Robots and Systems with Cognitive and Autonomy Capabilities*, (2010).
[23] J.L. Jimenez, I. Rano, and I. Minguez, 'Advances in the Framework for Automatic Evaluation of Obstacle Avoidance Methods', in *IROS Workshop on Benchmarks in Robotics Research*, (2007).
[24] Hiroaki Kitano, Minoru Asada, Yasuo Kuniyoshi, Itsuki Noda, and Eiichi Osawa, 'RoboCup: The Robot World Cup Initiative', in *IJCAI-95 workshop on entertainment and AI/ALife*, (1995).
[25] D Kraft, N Pugeault, E Baseski, M Popovic, D Kragic, S Kalkan, F Wörgötter, and N Krüger, 'Birth of the object: Detection of objectness and extraction of object shape through object action complexes', *International Journal of Humanoid Robotics*, **5**(2), 247–265, (2008).
[26] Raj Madhavan and Rolf Lakaemper, 'Benchmarking and Standardization of Intelligent Robotic Systems', *Intelligence*, (2009).
[27] *Performance Evaluation and Benchmarking of Intelligent Systems*, eds., Raj Madhavan, Edward Tunstel, and Elena Messina, Springer, 2009.
[28] O. Michel, 'Webots: Professional Mobile Robot Simulation', *International Journal of Advanced Robotic Systems*, **1**(1), 39–42, (2004).
[29] Olivier Michel, Fabien Rohrer, and Yvan Bourquin, 'Rat's Life: A Cognitive Robotics Benchmark', *European Robotics Symposium*, 223–232, (2008).
[30] Andrew Miller and Peter K. Allen, 'Graspit!: A Versatile Simulator for Robotic Grasping', *IEEE Robotics and Automation Magazine*, **11**(4), 110–122, (2004).
[31] Morgan Quigley, Ken Conley, Brian Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, and Andrew Y Ng, 'ROS: an open-source Robot Operating System', in *ICRA Workshop on Open Source Software*, (2009).
[32] Marc Raibert, Kevin Blankespoor, Gabriel Nelson, Rob Playter, and The BigDog Team, 'BigDog, the Rough-Terrain Quadruped Robot', in *Proceedings of the 17th World Congress of The International Federation of Automatic Control*, pp. 10822–10825, (2008).
[33] Ashutosh Saxena, Justin Driemeyer, and Andrew Y. Ng, 'Robotic Grasping of Novel Objects using Vision', *The International Journal of Robotics Research*, **27**(2), 157–173, (2008).
[34] Katherine M. Tsui, Munjal Desai, and Holly A. Yanco, 'Towards Measuring the Quality of Interaction: Communication through Telepresence Robots', in *Proceedings of the Performance Metrics for Intelligent Systems Workshop (PerMIS)*, (2012).
[35] Alan Turing, 'Computing Machinery and Intelligence', *Mind*, **59**, 433–60, (1950).
[36] S. Ulbrich, D. Kappler, T. Asfour, N. Vahrenkamp, A. Bierbaum, M. Przybylski, and R. Dillmann, 'The OpenGRASP Benchmarking Suite: An Environment for the Comparative Analysis of Grasping and Dexterous Manipulation', in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, (2011).
[37] Kai Zhou, Michael Zillich, and Markus Vincze, 'Mobile manipulation: Bring back the cereal box - Video proceedings of the 2011 CogX Spring School', in *8th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)*, pp. 873–873. Automation and Control Institute, Vienna University of Technology, 1040, Austria, IEEE, (2011).

Interactive Intelligence: Behaviour-based AI, Musical HCI and the Turing Test

Adam Linson, Chris Dobbyn and Robin Laney[1]

Abstract. The field of behaviour-based artificial intelligence (AI), with its roots in the robotics research of Rodney Brooks, is not predominantly tied to linguistic interaction in the sense of the classic Turing test (or, "imitation game"). Yet, it is worth noting, both are centred on a behavioural model of intelligence. Similarly, there is no intrinsic connection between musical AI and the language-based Turing test, though there have been many attempts to forge connections between them. Nonetheless, there are aspects of musical AI and the Turing test that can be considered in the context of non-language-based interactive environments–in particular, when dealing with *real-time* musical AI, especially interactive improvisation software. This paper draws out the threads of *intentional agency* and *human indistinguishability* from Turing's original 1950 characterisation of AI. On the basis of this distinction, it considers different approaches to musical AI. In doing so, it highlights possibilities for non-hierarchical interplay between human and computer agents.

1 Introduction

The field of behaviour-based artificial intelligence (AI), with its roots in the robotics research of Rodney Brooks, is not predominantly tied to linguistic interaction in the sense of the classic Turing test (or, "imitation game" [24]). Yet, it is worth noting, both are centred on a behavioural model of intelligence. Similarly, there is no intrinsic connection between musical AI and the language-based Turing test, though there have been many attempts to forge connections between them. The primary approach to applying the Turing test to music is in the guise of so-called "discrimination tests", in which human- and computer-generated musical output are compared (for an extensive critical overview of how the Turing test has been applied to music, see [1]). Nonetheless, there are aspects of musical AI and the Turing test that can be considered in the context of non-language-based interactive environments—in particular, when dealing with *real-time* musical AI, especially interactive improvisation software (see, for example, [23] and [8]). In this context, AI for non-hierarchical human-computer musical improvisation such as George Lewis' *Voyager* [16] and Turing's imitation game are both examples of "an open-ended and performative interplay between [human and computer] agents that are not capable of dominating each other" [21].

2 Background

It is useful here to give some context to the Turing test itself. In its original incarnation, the test was proposed as a thought experiment to explain the concept of a thinking machine to a public uninitiated in such matters [24]. Rather than as a litmus test of whether or not a machine could think (which is how the test is frequently understood), the test was in fact designed to help make sense of the concept of a machine that could think. Writing in 1950, he estimates "about fifty years' time" until the technology would be sufficient to pass a real version of the test and states his belief "that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted". Thus his original proposal remained a theoretical formulation: in principle, a machine could be invented with the capacity to be mistaken for a human; if this goal were accomplished, a reasonable person should accept the machine as a thinking entity. He is very clear about the behaviourist underpinnings of the experiment:

> May not machines carry out something which ought to be described as thinking but which is very different from what a man does? This objection is a very strong one, but at least we can say that if, nevertheless, a machine can be constructed to play the imitation game satisfactorily, we need not be troubled by this objection.

He goes on to describe the "imitation game" as one in which the machine should "try to provide answers that would naturally be given by a man". His ideas became the basis for what eventually emerged as the field of AI.

As Turing emphasised, the thought experiment consisted of an abstract, "imaginable" machine that—under certain conditions to ensure a level playing field—would be indistinguishable from a human, from the perspective of a human interrogator [24]. Presently, when the test is actually deployed in practice, it is easy to forget the essential role of the designer, especially given the fact that the computer "playing" the game is, to an extent, thrust into the spotlight. In a manner of speaking, the interactive computer takes centre stage, and attention is diverted from the underlying challenge set forth by Turing: *to determine the specifications* of the machine. Thus, one could say in addition to being a test for a given machine, it is also a creative design challenge to those responsible for the machine. The stress is on design rather than implementation, as Turing explicitly suggests imagining that any proposed machine functions perfectly according to its specifications (see [24], p. 449). If the creative design challenge were fulfilled, the computer would behave convincingly as a human, perhaps hesitating when appropriate and occasionally refusing to answer or giving incorrect answers such as the ones Turing imagines [24]:

> Q: Please write me a sonnet on the subject of the Forth Bridge.
> A: Count me out on this one. I never could write poetry.

[1] Faculty of Mathematics, Computing and Technology, Dept. of Computing, Open University, UK. Email: {a.linson, c.h.dobbyn, r.c.laney}@open.ac.uk

Q: Add 34957 to 70764.
A: (Pause about 30 seconds and then give as answer) 105621.

The implication of Turing's example is that the measure of success for those behind the machine lies in designing a system that is also as stubborn and fallible as humans, rather than servile and (theoretically) infallible, like an adding machine.

3 Two threads unraveled

Two threads can be drawn out of Turing's behavioural account of intelligence that directly pertain to contemporary AI systems: the first one concerns the kind of intentional agency suggested by his example answer, "count me out on this one"; the second one concerns the particular capacities and limitations of human embodiment, such as the human inability to perform certain calculations in a fraction of a second and the human potential for error. More generally, the second thread has to do with the broadly construed linguistic, social, mental and physical consequences of human physiology. Indeed, current theories of mind from a variety of disciplines provide a means for considering these threads separately. In particular, relevant investigations that address these two threads—described in this context as *intentional agency* and *human indistinguishability*—can be found in psychology, philosophy and cognitive science.

3.1 Intentional agency

The first thread concerns the notion of intentional agency, considered here separately from the thread of human indistinguishability. Empirical developmental psychology suggests that the human predisposition to attribute intentional agency to both humans *and* nonhumans appears to be present from infancy. Poulin-Dubois and Shultz chart childhood developmental stages over the first three years of life, from the initial ability to identify agency (distinguishing animate from inanimate objects) on to the informed attribution of intentionality, by inference of goal-directed behavior [22]. Csibra found that infants ascribed goal-directed behavior even to artificially animated inanimate objects, if the objects were secretly manipulated to display teleological actions such as obstacle avoidance [7]. Király, et al. identify the source of an infant's interpretation of a teleological action: "if the abstract cues of goal-directedness are present, even very young infants are able to attribute goals to the actions of a wide range of entities even if these are unfamiliar objects lacking human features" [10].

It is important to note that in the above studies, the infants were passive, remote observers, whereas the Turing test evaluates direct interaction. While the predisposition of infants suggests an important basis for such evaluation, more is needed to address interactivity. In another area of empirical psychology, a study of adults by Barrett and Johnson suggests that even a lack of apparent goals by a self-propelled (nonhuman) object can lead to the attribution of intentionality in an interactive context [2]. In particular, their test subjects used language normally reserved for humans and animals to describe the behaviour of artificially animated inanimate objects that appeared to exhibit resistance to direct control in the course of an interaction; when there was no resistance, they did not use such language. The authors of the study link the results of their controlled experiment to the anecdotal experience of the frustration that arises during interactions with artifacts such as computers or vehicles that "refuse" to cooperate. In other words, in an interactive context, too much passivity by an artificial agent may negate any sense of its apparent intentionality. This suggests that for an agent to remain apparently intentional during direct interaction, it must exhibit a degree of resistance along with the kind of adaptation to the environment that indicates its behaviour is being adjusted to attain a goal. These features appear to be accounted for in Turing's first example answer above: the answer is accommodating insofar as it is a direct response to the interrogator, but the show of resistance seems to enhance the sense of "intelligence". It is noteworthy that this particular thread, intentional agency, relates closely to Brooks' extension of intelligence to nonlinguistic, nonhuman intelligence, especially in relation to insect and other animal intelligence, which he has emulated in robotic form with his particular approach to AI (see [3]).

3.2 Human indistinguishability

The second thread, the idea that human capacities and limitations should be built into an AI system, strongly relates to many significant accounts of embodied, situated activity (see, for example, [9], [4] and [11]). These accounts focus on how the human body, brain, mind and environment fundamentally structure the process of cognition, which can be understood through observable behaviour. When dealing with AI, the focus on behaviour clearly ties back to Turing. These themes are also taken up in Brooks' behaviour-based AI approach, but, at least in his early research, he applies them primarily to nonhuman intelligence. In particular, he relates these themes to the kinds of adaptive behaviour described in the first thread. The differing properties of the second thread will come into sharper focus by returning to Turing's example, for a consideration of matters particular to humans.

Although Turing's example of pausing and giving an incorrect answer is a clear example of a human limitation over a machine, it is possible to give an inverted example of human and machine competence that applies equally well. If the question posed to the machine were instead "Is it easy to walk from here to the nearest supermarket?", the machine's answer would depend on how its designers handled the notion of "easy to walk to". In this case, the machine must not only emulate humans' abstract cognitive limitations when solving arithmetical problems; it must also be able to respond according to human bodily limitations. One could easily imagine a failed machine calculation: the supermarket is at the end of a single straight road, with no turns; it answers "yes, it is easy to walk to". But if the supermarket is very distant, or nearby but up a steep incline, then in order for the machine to give an answer that is indistinguishable from a human one, it must respond in a way that seems to share our embodied human limitations. Returning to the arithmetic example, as Doug Lenat points out, even some wrong answers are more human than others: "$93 - 25 = 78$ is more understandable than if the program pretends to get a wrong answer of 0 or -9998 for that subtraction problem" [14]. Although Lenat disputes the need for embodiment in AI (he prefers a central database of human common sense [13], which could likely address the "easy to walk to" example), it could be argued, following the above theoretical positions, that the set of humanlike wrong answers is ultimately determined by the "commonalities of our bodies and our bodily and social experience in the world" [11].

This second thread, which could also be characterised as *the attempt to seem humanlike*, is taken up in another nonlinguistic area of AI, namely, musical AI. Some "intelligent" computer music composition and performance systems appear very close to achieving human indistinguishability in some respects, although this is not always their explicitly stated purpose. For example, Manfred Clynes

describes a computer program that performs compositions by applying a single performer's manner of interpretation to previously unencountered material, across all instrumental voices [5]. He states that "our computer program plays music so that it is impossible to believe that no human performer is involved," which he qualifies by explaining the role of the human performer as a user of the software, who "instills the [musical performance] principles in the appropriate way". Taking an entirely different approach, David Cope, argues that a Turing-like test for creativity would be more appropriate to his work than a Turing test for intelligence [6]. On the other hand, he has called his well-known project "Experiments in Musical Intelligence" and he also makes reference to "intelligent music composition". Furthermore, he states that his system generates "convincing" music in the style of a given composer (by training the system with a corpus of human-composed music), and one can infer that, in this context, "convincing" at least approximates the notion of human indistinguishability. With a more critical articulation, Pearce and Wiggins carefully differentiate between a test for what Cope calls "convincing" and a Turing test for intelligence [19]. As they point out, despite the resemblance of the two approaches, testing for intelligence is distinct from determining the "(non-)membership of a machine composition in a set of human composed pieces of music". They also note the significant difference between an interactive test and one involving passive observation.

4 Broadening the interactive horizon

One reason for isolating these two threads is to recast Turing's ideas in a wider social context, one that is better attuned to the contemporary social understanding of the role of technology research: namely, that it is primarily intended (or even expected) to enhance our lives. Outside the thought experiment, in the realm of practical application, one might redirect the resources for developing a successful Turing test candidate (e.g., for the Loebner Prize) and instead apply them toward a different kind of interactive system. This proposed system could be built so that it might be easily identified as a machine (even if occasionally mistaken for a human), which seemingly runs counter to the spirit of the Turing test. However, with an altered emphasis, one could imagine the primary function of such a machine as *engaging* humans in a continuous process of interaction, for a variety of purposes, including (but not limited to) stimulating human creativity and providing a realm for aesthetic exploration.

One example of this kind of system is musical improvisation software that interacts with human performers in real time, in a mutually influential relationship between human and computer, such as Lewis' *Voyager*. In his software design, the interaction model strongly resembles the way in which Turing describes a computer's behaviour: it is responsive, yet it does not always give the expected answer, and it might interrupt the human interlocutor or steer the interaction in a different direction (see [16]). In the case of an interactive improvising music system, the environment in which the human and computer interact is not verbal conversation, but rather, a culturally specific aesthetic context for collaborative music-making. In this sense, a musical improvisation is not an interrogation in the manner presented by Turing, yet "test" conversations and musical improvisations are examples of free-ranging and open-ended human-computer interaction. Among other things, this kind of interaction can serve as a basis for philosophical enquiry and cognitive theory that is indeed very much in the spirit of Turing's 1950 paper [24] (see also [15] and [17]).

Adam Linson's *Odessa* is another intelligent musical system that is similarly rooted in freely improvised music (for a detailed description, see [18]). It borrows from Brooks' design approach in modelling the behaviour of an intentional agent, thus clearly taking up the first thread that has been drawn out here. Significantly, it isolates this thread (intentional agency) for study by abstaining from a direct implementation of many of the available methods for human emulation (aimed at the second thread), thus resulting in transparently nonhuman musical behaviour. Nonetheless, initial empirical studies suggest that the system affords an engaging and stimulating human-computer musical interaction. As the system architecture (based on Brooks' subsumption architecture) is highly extensible, future iterations of the system may add techniques for approximating fine-grained qualities of human musicianship. In the meantime, however, further studies are planned with the existing prototype, with the aim of providing insights into aspects of human cognition as well as intelligent musical agent design.

5 Conclusion

Ultimately, whether an interactive computer system is dealing with an interrogator in the imitation game or musically improvising with a human, the system must be designed to "respond in lived real time to unexpected, real-world input" [17]. This responsiveness takes the form of what sociologist Andrew Pickering calls the "dance of agency", in which a reciprocal interplay of resistance and accommodation produces unpredictable emergent results over time [20]. This description of a sustained, continuous play of forces that "interactively stablize" each other could be applied to freely improvised music, whether performed by humans exclusively, or by humans and computers together. Pickering points out a concept similar to the process of interactive stabilisation, 'heterogeneous engineering', elaborated in the work of his colleague John Law (see [12]); the latter, in its emphasis on productive output, is perhaps more appropriate to the musical context of free improvisation.

Although these theoretical characterisations may seem abstract, they concretely pertain to the present topic in that they seek to address the "open-ended and performative interplay between agents that are not capable of dominating each other" [21], where the agents may include various combinations of humans, computers and other entities, and the interplay may include linguistic, musical, physical and other forms of interaction. With particular relevance to the present context, Pickering applies his conceptual framework of agent interplay to the animal-like robots of Turing's contemporary, cybernetics pioneer Grey Walter, and those of Brooks, designed and built decades later [21]. Returning to the main theme, following Brooks, "the dynamics of the interaction of the robot and its environment are primary determinants of the structure of its intelligence" [3]. Thus, independent of its human resemblance, an agent's ability to negotiate with an unstructured and highly dynamic musical, social or physical environment can be treated as a measure of intelligence closely aligned with what Turing thought to be discoverable with his proposed test.

REFERENCES

[1] C. Ariza, 'The interrogator as critic: The turing test and the evaluation of generative music systems', *Computer Music Journal*, 33(2), 48–70, (2009).

[2] J.L. Barrett and A.H. Johnson, 'The role of control in attributing intentional agency to inanimate objects', *Journal of Cognition and Culture*, 3(3), 208–217, (2003).

[3] R.A. Brooks, *Cambrian intelligence: the early history of the new AI*, MIT Press, 1999.

[4] A. Clark, *Being There: Putting Brain, Body, and World Together Again*, MIT Press, 1997.
[5] M. Clynes, 'Generative principles of musical thought: Integration of microstructure with structure', *Communication and Cognition AI, Journal for the Integrated Study of Artificial Intelligence, Cognitive Science and Applied Epistemology*, 3(3), 185–223, (1986).
[6] D. Cope, *Computer Models of Musical Creativity*, MIT Press, 2005.
[7] G. Csibra, 'Goal attribution to inanimate agents by 6.5-month-old infants', *Cognition*, 107(2), 705–717, (2008).
[8] R.T. Dean, *Hyperimprovisation: Computer-interactive sound improvisation*, AR Editions, Inc., 2003.
[9] H. Hendriks-Jansen, *Catching ourselves in the act: Situated activity, interactive emergence, evolution, and human thought*, MIT Press, 1996.
[10] I. Király, B. Jovanovic, W. Prinz, G. Aschersleben, and G. Gergely, 'The early origins of goal attribution in infancy', *Consciousness and Cognition*, 12(4), 752–769, (2003).
[11] G. Lakoff and M. Johnson, *Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought*, Basic Books, 1999.
[12] J. Law, 'On the social explanation of technical change: The case of the portuguese maritime expansion', *Technology and Culture*, 28(2), 227–252, (1987).
[13] D.B. Lenat, 'Cyc: A large-scale investment in knowledge infrastructure', *Communications of the ACM*, 38(11), 33–38, (1995).
[14] D.B. Lenat, 'The voice of the turtle: Whatever happened to ai?', *AI Magazine*, 29(2), 11, (2008).
[15] G. Lewis, 'Interacting with latter-day musical automata', *Contemporary Music Review*, 18(3), 99–112, (1999).
[16] G. Lewis, 'Too many notes: Computers, complexity and culture in voyager', *Leonardo Music Journal*, 33–39, (2000).
[17] G. Lewis, 'Improvising tomorrow's bodies: The politics of transduction', *E-misférica*, 4.2, (2007).
[18] A. Linson, C. Dobbyn, and R. Laney, 'Improvisation without representation: artificial intelligence and music', in *Proceedings of Music, Mind, and Invention: Creativity at the Intersection of Music and Computation*, (2012).
[19] M. Pearce and G. Wiggins, 'Towards a framework for the evaluation of machine compositions', in *Proceedings of the AISB*, pp. 22–32, (2001).
[20] A. Pickering, *The mangle of practice: Time, agency, and science*, University of Chicago Press, 1995.
[21] A. Pickering, *The cybernetic brain: Sketches of another future*, University of Chicago Press, 2010.
[22] D. Poulin-Dubois and T.R. Shultz, 'The development of the understanding of human behavior: From agency to intentionality', in *Developing Theories of Mind*, eds., Janet W. Astington, Paul L. Harris, and David R. Olson, 109–125, Cambridge University Press, (1988).
[23] R. Rowe, *Machine musicianship*, MIT Press, 2001.
[24] A.M. Turing, 'Computing machinery and intelligence', *Mind*, 59(236), 433–460, (1950).

The ANYNT Project Intelligence Test Λ_{one}

Javier Insa-Cabrera[1] and **José Hernández-Orallo**[2] and **David L. Dowe**[3] and **Sergio España**[4] and **M.Victoria Hernández-Lloreda**[5]

Abstract. All tests in psychometrics, comparative psychology and cognition which have been put into practice lack a mathematical (computational) foundation or lack the capability to be applied to any kind of system (humans, non-human animals, machines, hybrids, collectives, etc.). In fact, most of them lack both things. In the past fifteen years, some efforts have been done to derive intelligence tests from formal intelligence definitions or vice versa, grounded on computational concepts. However, some of these approaches have not been able to create *universal* tests (i.e., tests which can evaluate any kind of subjects) and others have even failed to make a feasible test. The ANYNT project was conceived to explore the possibility of defining formal, universal and anytime intelligence tests, having a feasible implementation in mind. This paper presents the basics of the theory behind the ANYNT project and describes one of the test propotypes that were developed in the project: test Λ_{one}.

Keywords: (machine) intelligence evaluation, universal tests, artificial intelligence, Solomonoff-Kolmogorov complexity.

1 INTRODUCTION

There are many examples of intelligence tests which work in practice. For instance, in psychometrics and comparative psychology, tests are used to evaluate intelligence for a variety of subjects: children and adult Homo Sapiens, other apes, cetaceans, etc. In artificial intelligence, we are well aware of some incarnations and different variations of the Turing Test, such as the Loebner Prize or CAPTCHAs [32], which are also feasible and informative. However, they do not answer the pristine questions: what intelligence is and how it can be built.

In the past fifteen years, some efforts have been done to derive intelligence tests from formal intelligence definitions or vice versa, grounded on computational concepts. However, some of these approaches have not been able to create *universal* tests (i.e., tests which can evaluate any kind of subjects) and others have even failed to make a feasible test. The ANYNT project[6] was conceived to explore the possibility of defining formal, universal and anytime intelligence tests, having a feasible implementation in mind.

[1] DSIC, Universitat Politècnica de València, Spain. email: jinsa@dsic.upv.es
[2] DSIC, Universitat Politècnica de València, Spain. email: jorallo@dsic.upv.es
[3] Clayton School of Information Technology, Monash University, Australia. email: david.dowe@monash.edu
[4] PROS, Universitat Politècnica de València, Spain. email: sergio.espana@pros.upv.es
[5] Universidad Complutense de Madrid, Spain. email: vhlloreda@psi.ucm.es
[6] http://users.dsic.upv.es/proy/anynt/

In the ANYNT project we have been working on the design and implementation of a general intelligence test, which can be feasibly applied to a wide range of subjects. More precisely, the goal of the project is to develop intelligence tests that are: (1) *formal*, by using notions from Algorithmic Information Theory (a.k.a. Kolmogorov Complexity) [24]; (2) *universal*, so that they are able to evaluate the general intelligence of any kind of system (human, non-human animal, machine or hybrid). Each will have an appropriate interface that fits its needs; (3) *anytime*, so the more time is available for the evaluation, the more reliable the measurement will be.

2 BACKGROUND

In this section, we present a short introduction to the area of Algorithmic Information Theory and the notions of Kolmogorov complexity, universal distributions, Levin's Kt complexity, and its relation to the notions of compression, the Minimum Message Length (MML) principle, prediction, and inductive inference. Then, we will survey the approaches that have appeared using these formal notions in order to give mathematical definitions of intelligence or to develop intelligence tests from them, starting from the compression-enhanced Turing tests, the C-test, and Legg and Hutter's definition of Universal Intelligence.

2.1 Kolmogorov complexity and universal distributions

Algorithmic Information Theory is a field in computer science that properly relates the notions of computation and information. The key idea is the notion of the Kolmogorov Complexity of an object, which is defined as the length of the shortest program p that outputs a given string x over a machine U. Formally,

Definition 1 Kolmogorov Complexity

$$K_U(x) := \min_{p \text{ such that } U(p)=x} l(p)$$

where $l(p)$ denotes the length in bits of p and $U(p)$ denotes the result of executing p on U.

For instance, if $x = 1010101010101010$ and U is the programming language Lisp, then $K_{Lisp}(x)$ is the length in bits of the shortest program in Lisp that outputs the string x. The relevance of the choice of U depends mostly on the size of x. Since any universal machine can emulate another, it holds that for every two universal Turing machines U and V, there is a constant $c(U, V)$, which only depends on U and V and does not depend on x, such that for all x, $|K_U(x) - K_V(x)| \leq c(U, V)$. The value of $c(U, V)$ is relatively small for sufficiently long x.

From Definition 1, we can define the universal probability for machine U as follows:

Definition 2 Universal Distribution

Given a prefix-free machine[7] *U, the universal probability of string x is defined as:*

$$p_U(x) := 2^{-K_U(x)}$$

which gives higher probability to objects whose shortest description is small and gives lower probability to objects whose shortest description is large. Considering programs as hypotheses in the hypothesis language defined by the machine, paves the way for the mathematical theory of inductive inference and prediction. This theory was developed by Solomonoff [28], formalising Occam's razor in a proper way for prediction, by stating that the prediction maximising the universal probability will eventually discover any regularity in the data. This is related to the notion of Minimum Message Length for inductive inference [34][35][1][33] and is also related to the notion of data compression.

One of the main problems of Algorithmic Information Theory is that Kolmogorov Complexity is uncomputable. One popular solution to the problem of computability of K for finite strings is to use a time-bounded or weighted version of Kolmogorov complexity (and, hence, the universal distribution which is derived from it). One popular choice is Levin's Kt complexity [23][24]:

Definition 3 Levin's Kt Complexity

$$Kt_U(x) := \min_{p \text{ such that } U(p)=x} \{l(p) + \log time(U,p,x)\}$$

where $l(p)$ denotes the length in bits of p, $U(p)$ denotes the result of executing p on U, and $time(U,p,x)$ denotes the time[8] *that U takes executing p to produce x.*

Finally, despite the uncomputability of K and the computational complexity of its approximations, there have been some efforts to use Algorithmic Information Theory to devise optimal search or learning strategies. Levin (or universal) search [23] is an iterative search algorithm for solving inversion problems based on Kt, which has inspired other general agent policies such as Hutter's AIXI, an agent that is able to adapt optimally[9] in all environments where any other general purpose agent can be optimal [17], for which there is a working approximation [31][30].

2.2 Developing mathematical definitions and tests of intelligence

Following ideas from A.M. Turing, R.J. Solomonoff, E.M. Gold, C.S. Wallace, M. Blum, G. Chaitin and others, between 1997 and 1998 some works on enhancing or substituting the Turing Test [29] by inductive inference tests were developed, using Solomonoff prediction theory [28] and related notions, such as the Minimum Message Length (MML) principle. On the one hand, Dowe and Hajek [2][3][4] suggested the introduction of inductive inference problems in a somehow *induction-enhanced* or *compression-enhanced* Turing Test (they arguably called it non-behavioural) in order to, among other things, completely dismiss Searle's Chinese room [27] objection, and also because an inductive inference ability is a necessary (though possibly "not sufficient") requirement for intelligence.

Quite simultaneously and similarly, and also independently, in [13][6], intelligence was defined as the ability to comprehend, giving a formal definition of the notion of comprehension as the identification of a 'predominant' pattern from a given evidence, derived from Solomonoff prediction theory concepts, Kolmogorov complexity and Levin's Kt. The notion of comprehension was formalised by using the notion of "projectible" pattern, a pattern that has no exceptions (no noise), so being able to explain *every* symbol in the given sequence (and not only most of it).

From these definitions, the basic idea was to construct a *feasible* test as a set of series whose shortest pattern had no alternative projectible patterns of similar complexity. That means that the "explanation" of the series had to be much more plausible than other plausible hypotheses. The main objective was to reduce the subjectivity of the test — first, because we need to choose one reference universal machine from an infinite set of possibilities; secondly, because, even choosing one reference machine, two very different patterns could be consistent with the evidence and if both have similar complexities, their probabilities would be close, and choosing between them would make the series solution quite uncertain. With the constraints posed on patterns and series, both problems were not completely solved but minimised.

$k = 9$	:	a, d, g, j, ...	Answer: m
$k = 12$	:	a, a, z, c, y, e, x, ...	Answer: g
$k = 14$	:	c, a, b, d, b, c, c, e, c, d, ...	Answer: d

Figure 1. Examples of series of Kt complexity 9, 12, and 14 used in the C-test [6].

The definition was given as the result of a *test*, called C-test [13], formed by computationally-obtained series of increasing complexity. The sequences were formatted and presented in a quite similar way to psychometric tests (see Figure 1) and, as a result, the test was administered to humans, showing a high correlation with the results of a classical psychometric (IQ) test on the same individuals. Nonetheless, the main goal was that the test could eventually be administered to other kinds of intelligent beings and systems. This was planned to be done, but the work from [26] showed that machine learning programs could be specialised in such a way that they could score reasonably well on some of the typical IQ tests. A more extensive treatment of this phenomenon and the inadequacy of *current* IQ tests for evaluating machines can be found in [5]. This unexpected result confirmed that C-tests had important limitations and could not be considered universal in two ways, i.e., embracing the whole notion of intelligence, but perhaps only a part of it, and being applicable to any kind of subject (not only adult humans). The idea of extending these static tests to other factors or to make them interactive and extensible to other kinds of subjects by the use of rewards (as in the area of reinforcement learning) was suggested in [7][8], but not fully

[7] For a convenient definition of the universal probability, we need the requirement of U being a prefix-free machine (see, e.g., [24] for details). Note also that even for prefix-free machines there are infinitely many other inputs to U that will output x, so $p_U(x)$ is a strict lower bound on the probability that U will output x (given a random input)

[8] Here *time* does not refer to physical time but to computational time, i.e., computation steps taken by machine U. This is important, since the complexity of an object cannot depend on the speed of the machine where it is run.

[9] Optimality has to be understood in an asymptotic way. First, because AIXI is uncomputable (although resource-bounded variants have been introduced and shown to be optimal in terms of time and space costs). Second, because it is based on *a* universal probability over *a* machine, and this choice determines a constant term which may very important for small environments.

developed into actual tests. An illustration of the classical view of an environment in reinforcement learning is seen in Figure 2, where an agent can interact through actions, rewards and observations.

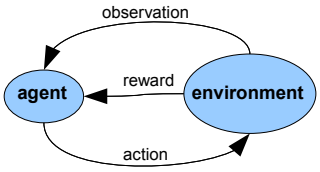


Figure 2. Interaction with an Environment.

A few years later, Legg and Hutter (e.g. [21],[22]) followed the previous steps and, strongly influenced by Hutter's theory of AIXI optimal agents [16], gave a new definition of machine intelligence, dubbed "Universal[10] Intelligence", also grounded in Kolmogorov complexity and Solomonoff's ("inductive inference" or) prediction theory. The key idea is that the intelligence of an agent is evaluated as some kind of sum (or weighted average) of performances in all the possible environments (as in Figure 2).

The definition based on the C-test can now be considered a static precursor of Legg and Hutter's work, where the environment outputs no rewards, and the agent is not allowed to make an action until several observations are seen (the inductive inference or prediction sequence). The point in favour of active environments (in contrast to passive environments) is that the former not only require inductive and predictive abilities to model the environment but also some planning abilities to effectively use this knowledge through actions. Additionally, perceptions, selective attention, and memory abilities must be fully developed. Not all this is needed to score well in a C-test, for instance.

While the C-test selects the problems by (intrinsic) difficulty (which can be chosen to fit the level of intelligence of the evaluee), Legg and Hutter's approach select problems by using a universal distribution, which gives more probability to simple environments. Legg and Hutter's definition, given an agent π, is given as:

Definition 4 Universal Intelligence [22]

$$\Upsilon(\pi, U) = \sum_{\mu=i}^{\infty} p_U(\mu) \cdot E\left(\sum_{i=1}^{\infty} r_i^{\mu,\pi}\right)$$

where μ is any environment coded on a universal machine U, with π being the agent to be evaluated, and $r_i^{\mu,\pi}$ the reward obtained by π in μ at interaction i. E is the expected reward on each environment, where environments are assigned with probability $p_U(\mu)$ using a universal distribution [28].

Definition 4, although very simple, captures one of the broadest definitions of intelligence: "the ability to adapt to a wide range of environments". However, this definition was not meant to be eventually converted into a test. In fact, there are three obvious problems in this definition regarding making it practical. First, we have two infinite sums in the definition: one is the sum over all environments, and the second is the sum over all possible actions (agent's life in each environment is infinite). And, finally, K is not computable. Additionally, we also have the dependence on the reference machine U. This dependence takes place even though we consider an infinite number of environments. The universal distribution for a machine U could give the higher probabilities $(0.5, 0.25, ...)$ to quite different environments than those given by another machine V.

Despite all these problems, it could seem that just making a random finite sample on environments, limiting the number of interactions or cycles of the agent with respect to the environment and using some computable variant of K, is sufficient to make it a practical test. However, on the one hand, this is not so easy, and, on the other hand, the definition has many other problems (some related and others unrelated).

The realisation of these problems and the search for solutions in the quest of a practical intelligence *test* is the goal of the ANYNT project.

3 ANYTIME UNIVERSAL TESTS

This section presents a summary of the theory in [11]. The reader is referred to this paper for further details.

3.1 On the difficulty of environments

The first issue concerns how to sample environments. Just using the universal distribution for this , as suggested by Legg and Hutter, will mean that very simple environments will be output again and again. Note that an environment μ with $K(\mu) = 1$ will appear half of the time. Of course, repeated environments must be ruled out, but a sample would almost become an enumeration from low to high K. This will still omit or underweight very complex environments because their probability is so low. Furthermore, measuring rewards on very small environments will get very unstable results and be very dependent on the reference machine. And even ignoring this, it is not clear that an agent that solves all the problems of complexity lower than 20 bits and none of those whose complexity is larger than 20 bits is more intelligent than another agent who does reasonably well on every environment.

This constrasts with the view of the C-test, which focus on the issue of difficulty and does not make the probability of a problem appearing inversely related to this difficulty. In any case, before going on, we need to clarify the notions of simple/easy and complex/difficult that are used here. For instance, just choosing an environment with high K does not ensure that the environment is indeed complex. As Figure 3 illustrates, the relation is unidirectional; given a low K, we can affirm that the environment will look simple. On the other hand, given an intuitively complex environment, K must be necessarily high.

Environment with high K $\Longleftarrow$ Intuitively complex (difficult) environment
Environment with low K $\Longrightarrow$ Intuitively simple (easy) environment

Figure 3. Relation between K and intuitive complexity.

Given this relation, only among environments with high K will we find complex environments, and, among the latter, not all of them will be difficult. From the agent's perspective, however, this is more extreme, since many environments with high K will contain difficult patterns that will never be accessed by the agent's interactions.

[10] The term 'universal' here does not refer to the definition (or a derived test) being applicable to any kind of agent, but to the use of Solomonoff's universal distribution and the view of the definition as an extremely general view of intelligence.

As a result, the environment will be *probabilistically* simple. Thus, giving most of the probability to environments with low K means that most of the intelligence measure will come from patterns that are extremely simple.

3.2 Selecting discriminative environments

Furthermore, many environments (either simple or complex) will be completely useless for evaluating intelligence, e.g., environments that stop interacting, environments with constant rewards, etc. If we are able to make a more accurate sample, we will be able to make a more efficient test procedure. The question here is to determine a non-arbitrary criterion to exclude some environments. For instance, Legg and Hutter's definition forces environments to interact infinitely, and since the description must be finite, there must be a pattern. This obviously includes environments such as "always output the same observation and reward". In fact, they are not only possible but highly probable on many reference machines. Another pathological case is an environment that "outputs observations and rewards at random". However, this has a high complexity if we assume deterministic environments. In both cases, the behaviour of any agent on these environments would almost be the same. In other words, they do not have *discriminative power*. Therefore, these environments would be useless for discriminating between agents.

In an interactive environment, a clear requirement for an environment to be discriminative is that what the agent does must have consequences on rewards. Thus, we will restrict environments to be sensitive to agents' actions. That means that a wrong action might lead the agent to part of the environment from which it can never return (non-ergodic), but at least the actions taken by the agent can modify the rewards in that subenvironment. More precisely, *we want an agent to be able to influence rewards at any point in any subenvironment*. This does not imply ergodicity but reward sensitivity at any moment. That means that we cannot reach a point from which rewards are given independently of what we do (a dead-end).

3.3 Symmetric rewards and balanced environments

An important issue is how to estimate rewards. If we only use positive rewards, we find some problems. For example, an increase in the score may originate from a really good behaviour on the environment or just because more rewards are accumulated since they are always positive. Instead, an average reward seems a better payoff function. Our proposal is to use symmetric rewards, which can range between -1 and 1:

Definition 5 Symmetric Rewards
We say an environment has symmetric rewards when:

$$\forall i : -1 \leq r_i \leq 1$$

If we set symmetric rewards, we also expect environments to be symmetric, or more precisely, to be balanced on how they give rewards. This can be seen in the following way. In a reliable test, we would like that many (if not all) environments give an expected 0 reward to random agents.

This excludes both hostile and benevolent environments, i.e., environments where doing randomly will get more negative (respectively positive) rewards than positive (respectively negative) rewards. In many cases it is not difficult to prove that a particular environment is balanced. Another approach is to set a reference machine that only generates balanced environments.

Using this approach on rewards, we can use an average to estimate the results on each environment, namely:

Definition 6 Average Reward
Given an environment μ, with n_i being the number of completed interactions, then the average reward for agent π is defined as follows:

$$v_\mu^\pi(n_i) = \frac{\sum_{i=1}^{n_i} r_i^{\mu,\pi}}{n_i}$$

Now we can calculate the expected value (although the limit may not exist) of the previous average, denoted by $E(v_\mu^\pi)$, for an arbitrarily large value of n_i.

To view the test framework in more detail, in [11] some of these issues (and many other problems) of the measure are solved. It uses a random finite sample of environments. It limits the number of interactions of the agent with respect to the environment. It selects a discriminative set of environments, etc.

4 ENVIRONMENT CLASS

The previous theory, however, does not make the choice for *an* environment class, but just sets some constraints on the kind of environments that can be used. Consequently, one major open problem is to make this choice, i.e., to find a proper (unbiased) environment class which follows the constraints and, more difficult, which can be feasibly implemented. Once this environment class is identified, we can use it to generate environments to run any of the tests variants. Additionally, it is not only necessary to determine the environment class, but also to determine the universal machine we will use to determine the Kolmogorov complexity of each environment, since the tests only use a (small) sample of environments, and the sample probability is defined in terms of the complexity.

In the previous section we defined a set of properties which are required for making environments discriminative, namely that observations and rewards must be sensitive to agent's actions and that environments are balanced. Given these constraints if we decide to generate environments without any constraint and then try to make a post-processing sieve to select which of them comply with all the constraints, we will have a computationally very expensive or even incomputable problem. So, the approach taken is to generate an environment class that ensures that these properties hold. In any case, we have to be very careful, because we would not like to restrict the reference machine to comply with these properties at the cost of losing their universality (i.e. their ability to emulate or include any computable function).

And finally, we would like the environment class to be user-friendly to the kind of systems we want to be evaluated (humans, non-human animals and machines), but without any bias in favour or against some of them.

According to all this, we define a universal environment class from which we can effectively generate valid environments, calculate their complexity and consequently derive their probability.

4.1 On actions, observations and space

Back to Figure 2 again, actions are limited by a finite set of symbols A, (e.g. $\{left, right, up, down\}$), rewards are taken from any subset R of rational numbers between -1 and 1, and observations are also

limited by a finite set O of possibilities (e.g., the contents of a grid of binary cells of $n \times m$, or a set of light-emitting diodes, LEDs). We will use a_i, r_i and o_i to (respectively) denote action, reward and observation at interaction i.

Apart from the behaviour of an environment, which may vary from very simple to very complex, we must first clarify the *interface*. How many actions are we going to allow? How many different observations? The very definition of environment makes actions a finite set of symbols and observations are also a finite set of symbols. It is clear that the minimum number of actions has to be two, but no upper limit seems to be decided a priori. The same happens with observations. Even choosing two for both, a sequence of interactions can be as rich as the expressiveness of a Turing machine.

Before getting into details with the interface, we have to think about environments that can contain agents. This is not only the case in real life (where agents are known as inanimate or animate objects, animals among the latter), but also a requirement for evolution and, hence, intelligence as we know it. The existence of several agents which can interact requires a *space*. The space is not necessarily a virtual or physical space, but also a set of common rules (or laws) that govern what the agents can perceive and what the agents can do. From this set of common rules, specific rules can be added to each agent. In the real world, this set of common rules is physics. All this has been extensively analysed in multi-agent systems (see e.g. [20] for a discussion).

The good thing about thinking of spaces is that a space entails the possible perceptions and actions. If we define a common space, we have many choices about observations and actions already taken.

A first (and common) idea for a space is a 2D grid. From a 2D grid, the observation is a picture of the grid with all the objects and agents inside. In a simple grid where we have agents and objects inside the cells, the typical actions are the movements left, right, up and down. Alternatively, of course, we could use a 3D space, since our world is 3D. In fact, there are some results using intelligence testing (for animals or humans) with a 3D interface [25][36].

The problem of a 2D or 3D grid is that it is clearly biased in favour of humans and many other animals which have hardwired abilities for orientation in this kind of spaces. Other kinds of animals or handicapped people (e.g. blind people) might have some difficulties in this type of spaces. Additionally, artificial intelligence agents would highly benefit by hardwired functionalities about Euclidean distance and 2D movement, without any real improvement in their general intelligence.

Instead we propose a more general kind of space. A 2D grid is a graph with a very special topology, where there are concepts which hold such as direction, adjacency, etc. A generalisation is a graph where the cells are freely connected to some other cells with no particular predefined pattern. This suggests a (generally) dimensionless space. Connections between cells would determine part or all the possible actions, and observations and rewards can be easily shown graphically.

4.2 Definition of the environment class

After the previous discussion, we are ready to give the definition of the environment class. First we must define the space and objects, and from here observations, actions and rewards. Before that, we have to define some constants that affect each environment. Namely, with $n_a = |A| \geq 2$ we denote the number of actions, with $n_c \geq 2$ the number of cells, and with n_ω the number of objects/agents (not including the agent which is to be evaluated and two special objects known as Good and Evil).

4.2.1 Space

The space is defined as a directed labelled graph of n_c nodes (or vertices), where each node represents a cell. Nodes are numbered, starting from 1, so cells are refered to as $C_1, C_2, \ldots, C_{n_c}$. From each cell we have n_a outgoing arrows (or arcs), each of them denoted as $C_i \rightarrow_\alpha C_j$, meaning that action $\alpha \in A$ goes from C_i to C_j. All the outgoing arrows from C_i are denoted by $\vec{C_i}$. At least two outgoing arrows cannot go to the same cell. Formally, $\forall C_i : \exists r_1, r_2 \in \vec{C_i}$ such that $r_1 = C_i \rightarrow_{\alpha_m} C_j$ and $r_2 = C_i \rightarrow_{\alpha_n} C_k$ with $C_j \neq C_k$ and $\alpha_m \neq \alpha_n$. At least one of the outgoing arrows from a cell must lead to itself (typically denoted by α_1 and is the first action). Formally, $\forall C_i : \exists r \in \vec{C_i}$ such that $r = C_i \rightarrow_{\alpha_1} C_i$.

A path from C_i to C_m is a sequence of arrows $C_i \rightarrow C_j, C_j \rightarrow C_k, \ldots, C_l \rightarrow C_m$. The graph must be strongly connected, i.e., all cells must be connected (i.e. there must be a walk over the graph that goes through all its nodes), or, in other words, for every two cells C_i, C_j there exists a path from C_i to C_j.

4.2.2 Objects

Cells can contain objects from a set of predefined objects Ω, with $n_\omega = |\Omega|$. Objects, denoted by ω_i can be animate or inanimate, but this can only be perceived by the rules each object has. An object is inanimate (for a period or indefinitely) when it performs action α_1 repeatedly. Objects can perform actions following the space rules, but apart from these rules, they can have any behaviour, either deterministic or not. Objects can be reactive and can be defined to act with different actions according to their observations. Objects perform one and only one action at each interaction of the environment (except from the special objects Good and Evil, which can perform several actions in a row).

Apart from the evaluated agent π, as we have mentioned, there are two special objects called Good and Evil. Good and Evil must have the same behaviour. By the *same* behavior we do not mean that they perform the same movements, but they have the same *logic* or *program* behind them.

Objects can share a same cell, except Good and Evil, which cannot be at the same cell. If their behaviour leads them to the same cell, then one (chosen randomly with equal probability) moves to the intended cell and the other remains at its original cell. Because of this, the environment becomes stochastic (non-deterministic).

Objects are placed randomly at the cells with the initialisation of the environment. This is another source of stochastic behaviour.

4.2.3 Observations and Actions

The observation is a sequence of cell contents. The cells are ordered by their number. Each element in the sequence shows the presence or absence of each object, included the evaluated agent. Additionally, each cell which is reachable by an action includes the information of that action leading to the cell.

4.2.4 Rewards

Raw rewards are defined as a function of the position of the evaluated agent π and the positions of Good and Evil.

For the rewards, we will work with the notion of trace and the notion of "cell reward", that we denote by $r(C_i)$. Initially, $r(C_i) = 0$

for all i. Cell rewards are updated by the movements of Good and Evil. At each interaction, we set r_i^{Good} to the cell reward where Good is and $-r_i^{Evil}$ to the cell reward where Evil is. Each interaction, all the other cell rewards are divided by a constant n (for example $n = 2$). So, an intuitive way of seeing this is that Good leaves a positive trace and Evil leaves a negative trace. The agent π *eats* the rewards it finds in the cells it occupies. We mean it *eats*, since just after getting the reward, the cell reward is set to 0. Note that if $n = \infty$, then Good and Evil do not leave any trace of rewards.

When π moves to a cell, it gets the cell reward which is at that cell, i.e. the accumulated reward $\rho = \rho + r(C_i)$. To calculate the average of the rewards, we divide the accumulated reward by the final number of interactions (denoted by n_i). The justification for this option is further investigated in [10].

4.2.5 Properties

All the properties mentioned in the previous section (observation-sensitiveness, reward-sensitiveness and balancedness), are met by the environment class described here. For a proof of these properties for this environment class see [9].

5 DESCRIPTION OF THE TEST Λ_{one}

In this section we will explain how an actual test is constructed. In particular, we will see one of our prototypes: Λ_{one}. We will explain how exercises are arranged, we will see an interface for humans and we will comment on some experimental results with this test.

5.1 Episodes

Tests are sequence of exercises (or environments). In particular, Λ_{one}uses 7 environments, each with a number of cells (n_c) from 3 to 9. The size of the patterns for Good and Evil is made proportional to the number of cells, using n_c actions (on average). In each environment, we allow $10 \times (n_c - 1)$ steps, so the agents have the chance to detect any pattern in the environment (exploration) and also have some further steps to exploit the findings (in case a pattern is actually there). The limitation of the number of environments and steps is justified because the tests are meant to be applied to biological agents in a reasonable period of time (e.g., 20 minutes) and we estimate an average of 4 seconds per action. Table 1 shows the choices we made for the test:

Env. #	No. cells (n_c)	No. steps (m)	Pattern length (on average)
1	3	20	3
2	4	30	4
3	5	40	5
4	6	50	6
5	7	60	7
6	8	70	8
7	9	80	9
TOTAL	-	350	-

Table 1. Setting for the 7 environments which compose Λ_{one}.

Before each exercise starts, a random environment is created (by generating the space topology and the behaviour of Good and Evil) using the environment distribution, and the three agents (Good, Evil and the evaluated agent) are placed into the generated space. The interaction starts when the evaluated agent decides which cell to move to. Then, the three agents are moved simultaneously. Once Good and Evil move to a cell, they leave their rewards in their respective cells, and the rest of the cell rewards are deleted. Finally, the evaluated agent collects the reward of the cell where it is, and a new interaction is started. When the test ends, the score of the evaluated agent is calculated as the average of the collected rewards over the whole exercise.

5.2 Interfaces

Applying a test to a kind of individual requires an *interface* for that kind. Clearly, the same test may require different interfaces for adult humans, blind people, a dolphin or a machine. Given the formal definition of the environments, it is relatively easy to figure out an interface for machines. In fact, in our case, we just connect the environments to reinforcement learning agents, in the traditional way.

For biological agents, constructing a user interface is a delicate issue, in the sense that we must ensure that no designing decision should contaminate the evaluation. The interface for humans has been designed with the following principles in mind: 1) The subject must not perceive anything that could distract it from the test. 2) The signs used to represent observations should not have an implicit meaning for the subject (e.g. no skull-and-bones for the Evil agent). 3) Actions and rewards should be easily interpreted by the subject to avoid a cognitive overhead.

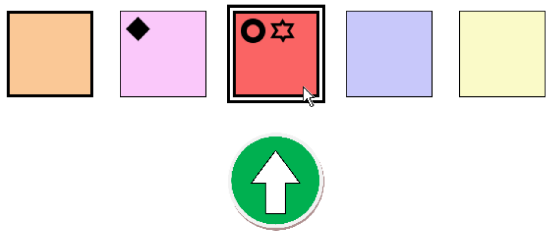


Figure 4. A snapshot of an exercise using a human interface.

In Fig. 4 we can see a snapshot of one exercise using the human interface. In the figure we see an environment with five cells. In the third cell we can see the evaluated agent (◎) sharing the cell with Good (✡). Evil (◇) can be seen in the second cell. The cells directly accessible by the evaluated agent are denoted by thick edges. The third cell has a double border because it is accessible and the user has the cursor over it. In this interaction the evaluated agent has received a positive reward (represented by a green circle with an up arrow) because it has coincided with Good in the cell.

5.3 Experiments

During the project we have made some experiments to analyse how the test works. We just include a brief excerpt from some of them. In [19] we evaluated the performance of a reinforcement learning algorithm. For this experiment, we analysed the results of a well known algorithm in reinforcement learning known as QLearning [37]. For the evaluation, we let QLearning interact with several environment

complexities, and we analysed whether the obtained results correlated with the measure of difficulty. The results were clear, showing that the evaluation obtains the expected results in terms of the relation between expected reward and theoretical problem difficulty. Also, it showed reasonable differences with other baseline algorithms (e.g. a random algorithm). All this supported the idea that the test and the environment class used are on the right direction for evaluating a specific kind of system. However, the main question was whether the approach was in the right direction in terms of constructing universal tests. In other words, it was still necessary to demonstrate if the test serves to evaluate several kinds of systems and put their results on the same scale.

In [18] we compared the results of two different systems (humans and AI algorithms), by using the prototype described in this paper and the interface for humans. We set both systems to interact with exactly the same environments. The results, not surprisingly, did not show the expected difference in intelligence between reinforcement learning algorithms and humans. This is explained by several reasons. One of them is that the environments were still relatively simple and reinforcement learning algorithms could still capture and represent all the state matrix for these problems with some partial success. Another reason is that exercises were independent, so humans could not reuse what they were learning on some exercises for others, an issue where humans are supposed to be better than these simple reinforcement algorithms. Also, another possibility is the fact that the environments had very few agents and the few agents that existed were not reactive. This makes the state space bounded, which is beneficial for Q-learning. Similarly, the environments had no noise. All these decisions were made on purpose to keep things simple and also to be able to formally derive the complexity of the environments. In general, other explanations can be found as well, since the lack of other interactive agents can be seen as a lack of social behaviours, as we explored in subsequent works [12].

Of course, test $\Lambda_{one}$was just a first prototype which does not incorporate many of the features of an anytime intelligence test and the measuring framework. Namely, the prototype is not anytime, so the test does not adapt its complexity to the subject that is evaluating. Also, we made some simplifications to the environment class, causing objects to lose reactivity. Furthermore, it is very difficult to construct any kind of social behaviour by creating agents from scratch. These and other issues are being addressed in new prototypes, some of them under development.

6 CONCLUDING REMARKS

The ANYNT project aimed at exploring the possibility of formal, universal and feasible tests. As already said, test $\Lambda_{one}$is just one prototype that does not implement all the features of the theory of *anytime universal tests*. However, it is already very informative. For instance, the experimental results show that the test Λ_{one}goes in the right direction, but it still fails to capture some components of intelligence that should put different kinds of individuals on the right scale.

In defence of test Λ_{one}, we have to say that it is quite rare in the literature to find the same test applied to different kinds of individuals[11]. In fact, as argued in [5], relatively simple programs can get good scores on conventional IQ tests, while small children (with high potential intelligence) will surely fail. Similarly, illiterate people and most children would score very badly at the Turing Test, for instance. And humans are starting to struggle with many CAPTCHAs.

All this means that many feasible and practical tests work because they are specialised for specific populations. As long as the diversity of subjects is enlarged, measuring intelligence becomes more difficult and less accurate. As a result, the mere possibility of constructing universal tests is still a hot question. While many may think that this is irresoluble, we think that unless an answer to this question is found, it will be very difficult (if not impossible) to assess the diversity of intelligent agents that are envisaged for the forthcoming decades. Being one way or another, there is clearly an ocean of scientific questions beyond the Turing Test.

[11] The only remarkable exceptions are the works in comparative psychology, such as [14][15], which are conscious of the difficulties of using the same test, with different interfaces, for different subjects.

ACKNOWLEDGEMENTS

This work was supported by the MEC projects EXPLORA-INGENIO TIN 2009-06078-E, CONSOLIDER-INGENIO 26706 and TIN 2010-21062-C02-02, and GVA project PROMETEO/2008/051. Javier Insa-Cabrera was sponsored by Spanish MEC-FPU grant AP2010-4389.

REFERENCES

[1] D. L. Dowe, 'Foreword re C.S. Wallace', *The Computer Journal*, **51**(5), 523–560, Christopher Stewart WALLACE (1933–2004) memorial special issue, (2008).

[2] D. L. Dowe and A. R. Hajek, 'A computational extension to the Turing Test', *in Proceedings of the 4th Conference of the Australasian Cognitive Science Society, University of Newcastle, NSW, Australia*, (1997).

[3] D. L. Dowe and A. R. Hajek, 'A computational extension to the Turing Test', *Technical Report #97/322, Dept Computer Science, Monash University, Melbourne, Australia, 9pp, http://www.csse.monash.edu.au/publications/1997/tr-cs97-322-abs.html*, (1997).

[4] D. L. Dowe and A. R. Hajek, 'A non-behavioural, computational extension to the Turing Test', in *International conference on computational intelligence & multimedia applications (ICCIMA'98), Gippsland, Australia*, pp. 101–106, (1998).

[5] D. L. Dowe and J. Hernandez-Orallo, 'IQ tests are not for machines, yet', *Intelligence*, **40**(2), 77–81, (2012).

[6] J. Hernández-Orallo, 'Beyond the Turing Test', *Journal of Logic, Language and Information*, **9**(4), 447–466, (2000).

[7] J. Hernández-Orallo, 'Constructive reinforcement learning', *International Journal of Intelligent Systems*, **15**(3), 241–264, (2000).

[8] J. Hernández-Orallo, 'On the computational measurement of intelligence factors', in *Performance metrics for intelligent systems workshop*, ed., A. Meystel, pp. 1–8. National Institute of Standards and Technology, Gaithersburg, MD, U.S.A., (2000).

[9] J. Hernández-Orallo, 'A (hopefully) non-biased universal environment class for measuring intelligence of biological and artificial systems', in *Artificial General Intelligence, 3rd International Conference AGI, Proceedings*, eds., Marcus Hutter, Eric Baum, and Emanuel Kitzelmann, "Advances in Intelligent Systems Research" series, pp. 182–183. Atlantis Press, (2010).

[10] J. Hernández-Orallo, 'On evaluating agent performance in a fixed period of time', in *Artificial General Intelligence, 3rd Intl Conf*, ed., M. Hutter et al., pp. 25–30. Atlantis Press, (2010).

[11] J. Hernández-Orallo and D. L. Dowe, 'Measuring universal intelligence: Towards an anytime intelligence test', *Artificial Intelligence*, **174**(18), 1508–1539, (2010).

[12] J. Hernández-Orallo, D. L. Dowe, S. España-Cubillo, M. V. Hernández-Lloreda, and J. Insa-Cabrera, 'On more realistic environment distributions for defining, evaluating and developing intelligence', in *Artificial General Intelligence 2011*, eds., J. Schmidhuber, K.R. Thórisson, and M. Looks (eds), volume 6830 of *LNAI*, pp. 82–91. Springer, (2011).

[13] J. Hernández-Orallo and N. Minaya-Collado, 'A formal definition of intelligence based on an intensional variant of kolmogorov complexity', in *In Proceedings of the International Symposium of Engineering of Intelligent Systems (EIS'98)*, pp. 146–163. ICSC Press, (1998).

[14] E. Herrmann, J. Call, M. V. Hernández-Lloreda, B. Hare, and M. Tomasell, 'Humans have evolved specialized skills of social cognition: The cultural intelligence hypothesis', *Science*, **Vol 317**(5843), 1360–1366, (2007).
[15] E. Herrmann, M. V. Hernández-Lloreda, J. Call, B. Hare, and M. Tomasello, 'The structure of individual differences in the cognitive abilities of children and chimpanzees', *Psychological Science*, **21**(1), 102, (2010).
[16] M. Hutter, *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*, Springer, 2005.
[17] M. Hutter, 'Universal algorithmic intelligence: A mathematical top→down approach', in *Artificial General Intelligence*, eds., B. Goertzel and C. Pennachin, Cognitive Technologies, 227–290, Springer, Berlin, (2007).
[18] J. Insa-Cabrera, D. L. Dowe, S. España-Cubillo, M. V. Hernández-Lloreda, and J. Hernández-Orallo, 'Comparing humans and AI agents', in *Artificial General Intelligence 2011*, eds., J. Schmidhuber, K.R. Thórisson, and M. Looks (eds), volume 6830 of *LNAI*, pp. 122–132. Springer, (2011).
[19] J. Insa-Cabrera, D. L. Dowe, and J. Hernández-Orallo, 'Evaluating a reinforcement learning algorithm with a general intelligence test', in *CAEPIA, Advances in Artificial Intelligence*, volume 7023 of *LNCS*, pp. 1–11. Springer, (2011).
[20] D. Keil and D. Goldin, 'Indirect interaction in environments for multi-agent systems', *Environments for Multi-Agent Systems II*, 68–87, (2006).
[21] S. Legg and M. Hutter, 'A universal measure of intelligence for artificial agents', in *International Joint Conference on Artificial Intelligence*, volume 19, p. 1509, (2005).
[22] S. Legg and M. Hutter, 'Universal intelligence: A definition of machine intelligence', *Minds and Machines*, **17**(4), 391–444, (2007). http://www.vetta.org/documents/UniversalIntelligence.pdf.
[23] L. A. Levin, 'Universal sequential search problems', *Problems of Information Transmission*, **9**(3), 265–266, (1973).
[24] M. Li and P. Vitányi, *An introduction to Kolmogorov complexity and its applications (3rd ed.)*, Springer-Verlag New York, Inc., 2008.
[25] F. Neumann, A. Reichenberger, and M. Ziegler, 'Variations of the turing test in the age of internet and virtual reality', in *Proceedings of the 32nd annual German conference on Advances in artificial intelligence*, pp. 355–362. Springer-Verlag, (2009).
[26] P. Sanghi and D. L. Dowe, 'A computer program capable of passing IQ tests', in *Proc. 4th ICCS International Conference on Cognitive Science (ICCS'03), Sydney, Australia*, pp. 570–575, (July 2003).
[27] J. Searle, 'Minds, brains, and programs', *Behavioral and Brain Sciences*, **3**(3), 417–457, (1980).
[28] R. J. Solomonoff, 'A formal theory of inductive inference. Part I', *Information and control*, **7**(1), 1–22, (1964).
[29] A. M. Turing, 'Computing machinery and intelligence', *Mind*, **59**, 433–460, (1950).
[30] J. Veness, K. S. Ng, M. Hutter, and D. Silver, 'Reinforcement learning via AIXI approximation', in *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI-10)*, pp. 605–611, (2010).
[31] J. Veness, K.S. Ng, M. Hutter, W. Uther, and D. Silver, 'A Monte Carlo AIXI Approximation', *Journal of Artificial Intelligence Research*, **40**(1), 95–142, (2011).
[32] L. Von Ahn, M. Blum, and J. Langford, 'Telling humans and computers apart automatically', *Communications of the ACM*, **47**(2), 56–60, (2004).
[33] C. S. Wallace, *Statistical and Inductive Inference by Minimum Message Length*, Ed. Springer-Verlag, 2005.
[34] C. S. Wallace and D. M. Boulton, 'A information measure for classification', *The Computer Journal*, **11**(2), 185–194, (1968).
[35] C. S. Wallace and D. L. Dowe, 'Minimum message length and Kolmogorov complexity', *Computer Journal*, **42**(4), 270–283, (1999). Special issue on Kolmogorov complexity.
[36] D.A. Washburn and R.S. Astur, 'Exploration of virtual mazes by rhesus monkeys (macaca mulatta)', *Animal Cognition*, **6**(3), 161–168, (2003).
[37] C.J.C.H. Watkins and P. Dayan, 'Q-learning', *Mach. learning*, **8**(3), 279–292, (1992).

Turing Machines and Recursive Turing Tests

José Hernández-Orallo[1] and **Javier Insa-Cabrera**[2] and **David L. Dowe**[3] and **Bill Hibbard**[4]

Abstract. The Turing Test, in its standard interpretation, has been dismissed by many as a practical intelligence test. In fact, it is questionable that the imitation *game* was meant by Turing himself to be used as a *test* for evaluating machines and measuring the progress of artificial intelligence. In the past fifteen years or so, an alternative approach to measuring machine intelligence has been consolidating. The key concept for this alternative approach is not the Turing *Test*, but the Turing *machine*, and some theories derived upon it, such as Solomonoff's theory of prediction, the MML principle, Kolmogorov complexity and algorithmic information theory. This presents an antagonistic view to the Turing test, where intelligence tests are based on formal principles, are not anthropocentric, are meaningful computationally and the abilities (or factors) which are evaluated can be recognised and quantified. Recently, however, this computational view has been touching upon issues which are somewhat related to the Turing Test, namely that we may need other intelligent agents in the tests. Motivated by these issues (and others), this paper links these two antagonistic views by bringing some of the ideas around the Turing *Test* to the realm of Turing *machines*.

Keywords: Turing Test, Turing machines, intelligence, learning, imitation games, Solomonoff-Kolmogorov complexity.

1 INTRODUCTION

Humans have been evaluated by other humans in all periods of history. It was only in the 20th century, however, that psychometrics was established as a scientific discipline. *Other* animals have also been evaluated by humans, but certainly not in the context of psychometric tests. Instead, comparative cognition is nowadays an important area of research where non-human animals are evaluated and compared. *Machines* —yet again differently— have also been evaluated by humans. However, no scientific discipline has been established for this.

The Turing Test [31] is still the most popular test for machine intelligence, at least for philosophical and scientific discussions. The Turing Test, as a measurement *instrument* and not as a philosophical argument, is very different to the instruments other disciplines use to measure intelligence in a scientific way. The Turing Test resembles a much more customary (and non-scientific) assessment, which happens when humans interview or evaluate other humans (for whatever reason, including, e.g., personnel selection, sports[1] or other competitions). The most relevant (and controversial) feature of the Turing Test is that it takes *humans* as a touchstone to which machines should be compared. In fact, the comparison is not performed by an objective criterion, but assessed by *human* judges, which is not without controversy. Another remarkable feature (and perhaps less controversial) is that the Turing Test is set on an intentionally restrictive interaction channel: a teletype conversation. Finally, there are some features about the Turing Test which make it more general than other kinds of intelligence tests. For instance, it is becoming increasingly better known that programs can do well at human IQ tests [32][8], because ordinary IQ tests only evaluate narrow abilities and assume that narrow abilities accurately reflect human abilities across a broad set of tasks, which may not hold for non-human populations. The Turing test (and some formal intelligence measures we will review in the following section) can test broad sets of tasks.

We must say that Turing cannot be blamed for all the controversy. The purpose of Turing's imitation game [37] was to show that intelligence could be assessed and recognised in a behavioural way, without the need for directly measuring or recognising some other physical or mental issues such as thinking, consciousness, etc. In Turing's view, intelligence can be just seen as a cognitive ability (or property) that some machines might have and others might not. In fact, the standard scientific view should converge to defining intelligence as an ability that some systems: humans, non-human animals, machines —and collectives thereof—, might or might not have, or, more precisely, might have to a larger or lesser degree. This view has clearly been spread by the popularity of psychometrics and IQ tests.[2]

While there have been many variants and extensions of the Turing Test (see [33] or [31] for an account of these), none of them (and none of the approaches in psychometrics and animal cognition, either) have provided a formal, mathematical definition of what in-

[1] DSIC, Universitat Politècnica de València, Spain. email: jorallo@dsic.upv.es
[2] DSIC, Universitat Politècnica de València, Spain. email: jinsa@dsic.upv.es
[3] Clayton School of Information Technology, Monash University, Australia. email: david.dowe@monash.edu
[4] Space Science and Engineering Center, University of Wisconsin - Madison, USA. email: test@ssec.wisc.edu

[1] In many sports, to see how good a player is, we want competent judges but also appropriate team-mates and opponents. Good tournaments and competitions are largely designed so as to return (near) maximal expected information.

[2] In fact, the notion of consciousness and other phenomena is today better separated from intelligence than it was sixty years ago. They are now seen as related but different things. For instance, nobody doubts that a team of people can score well in a single IQ test (working together). In fact, the team, using a teletype communication as in the Turing Test, can dialogue, write poetry, make jokes, do complex mathematics and all these human things. They can even do these things continuously for days or weeks, while some of the particular individuals rest, eat, go to sleep, die, etc. Despite all of this happening *on the other side of the teletype communication*, the system is just regarded as one subject. So the fact that we can effectively measure the cognitive abilities of the team or even make the team pass the Turing Test does not lead us directly to statements such as 'the team has a mind' or 'the team is conscious'. At most, we say this in a figurative sense, as we use it for the *collective consciousness* of a company or country. In the end, the 'team of people' is one of the best arguments against Searle's Chinese room and a good reference whenever we are thinking about evaluating intelligence.

telligence is and how it can be measured.

A different approach is based on one of the things that the Turing Test is usually criticised for: *learning*[3]. This alternative approach requires a proper definition of learning, and actual mechanisms for measuring learning ability. Interestingly, the answer to this is given by notions devised from Turing machines. In the 1960s, Ray Solomonoff 'solved' the problem of induction (and the related problems of prediction and learning) [36] by the use of Turing machines. This, jointly with the theory of inductive inference given by the Minimum Message Length (MML) principle [39, 40, 38, 5], algorithmic information theory [1], Kolmogorov complexity [25, 36] and compression theory, paved the way in the 1990s for a new approach for defining and measuring intelligence based on algorithmic information theory. This approach will be summarised in the next section.

While initially there was some connection to the Turing Test, this line of research has been evolving and consolidating in the past fifteen years (or more), cutting all the links to the Turing Test. This has provided important insights into what intelligence is and how it can be measured, and has given clues to the (re-)understanding of other areas where intelligence is defined and measured, such as psychometrics and animal cognition.

An important milestone of this journey has been the recent realisation in this context that (social) intelligence is the ability to perform well in an environment full of other agents of similar intelligence. This is a consequence of some experiments which show that when performance is measured in environments where no other agents coexist, some important traits of intelligence are not fully recognised. A solution for this has been formalised as the so-called Darwin-Wallace distribution of environments (or tasks) [18]. The outcome of all this is that it is increasingly an issue whether intelligence might be needed to measure intelligence. But this is not because we might need intelligent judges as in the Turing Test, but because we may need other intelligent agents to become part of the exercises or tasks an intelligence test should contain (as per footnote 1).

This seems to take us back to the Turing Test, a point some of us deliberately abandoned more than fifteen years ago. Re-visiting the Turing Test now is necessarily very different, because of the technical companions, knowledge and results we have gathered during this journey (universal Turing machines, compression, universal distributions, Solomonoff-Kolmogorov complexity, MML, reinforcement learning, etc.).

The paper is organised as follows. Section 2 introduces a short account of the past fifteen years concerning definitions and tests of machine intelligence based on (algorithmic) information theory. It also discusses some of the most recent outcomes and positions in this line, which have led to the notion of Darwin-Wallace distribution and the need for including other intelligent agents in the tests, suggesting an inductive (or recursive, or iterative) test construction and definition. This is linked to the notion of recursive Turing Test (see [32, sec. 5.1] for a first discussion on this). Section 3 analyses the base case by proposing several schemata for evaluating systems that are able to imitate Turing machines. Section 4 defines different ways of doing the recursive step, inspired by the Darwin-Wallace distribution and ideas for making this feasible. Section 5 briefly explores how all this might develop, and touches upon concepts such as universality in Turing machines and potential intelligence, as well as some suggestions as to how machine intelligence measurement might develop in the future.

[3] This can be taken as further evidence for Turing not conceiving the imitation test as an actual test for intelligence, because the issue about machines being able to learn was seen as inherent to intelligence for Turing [37, section 7], and yet the Turing Test is not especially good at detecting learning ability *during* the test.

2 MACHINE INTELLIGENCE MEASUREMENT USING TURING MACHINES

There are, of course, many proposals for intelligence definitions and tests *for machines* which are not based on the Turing Test. Some of them are related to psychometrics, some others may be related to other areas of cognitive science (including animal cognition) and some others originate from artificial intelligence (e.g., some competitions running on specific tasks such as planning, robotics, games, reinforcement learning, ...). For an account of some of these, the reader can find a good survey in [26]. In this section, we will focus on approaches which use Turing machines (and hence computation) as a basic component for the definition of intelligence and the derivation of tests for machine intelligence.

Most of the views of intelligence in computer science are sustained over a notion of intelligence as a special kind of information processing. The nature of information, its actual content and the way in which patterns and structure can appear in it can only be explained in terms of algorithmic information theory. The Minimum Message Length (MML) principle [39, 40] and Solomonoff-Kolmogorov complexity [36, 25] capture the intuitive notion that there is structure –or redundancy– in data if and only if it is compressible, with the relationship between MML and (two-part) Kolmogorov complexity articulated in [40][38, chap. 2][5, sec. 6]. While Kolmogorov [25] and Chaitin [1] were more concerned with the notions of randomness and the implications of all this in mathematics and computer science, Solomonoff [36] and Wallace [39] developed the theory with the aim of explaining how learning, prediction and inductive inference work. In fact, Solomonoff is said to have 'solved' the problem of induction [36] by the use of Turing machines. He was also the first to introduce the notions of universal distribution (as the distribution of strings given by a UTM from random input) and the invariance theorem (which states that the Kolmogorov complexity of a string calculated with two different reference machines only differs by a constant which is independent of the string).

Chaitin briefly made mention in 1982 of the potential relationship between algorithmic information theory and measuring intelligence [2], but actual proposals in this line did not start until the late 1990s. The first proposal was precisely introduced over a Turing Test and as a response to Searle's Chinese room [35], where the subject was *forced* to learn. This *induction-enhanced* Turing Test [7][6] could then evaluate a general inductive ability. The importance was not that any kind of ability could be included in the Turing Test, but that this ability could be formalised in terms of MML and related ideas, such as (two-part) compression.

Independently and near-simultaneously, a new intelligence test (C-test) [19] [12] was derived as sequence prediction problems which were generated by a universal distribution [36]. The difficulty of the exercises was mathematically derived from a variant of Kolmogorov complexity, and only exercises with a certain degree of difficulty were included and weighted accordingly. These exercises were very similar to those found in some IQ tests, but here they were created from computational principles. This work 'solved' the traditional subjectivity objection of the items in IQ tests, i.e., since the continuation of each sequence was derived from its shortest explanation. However, this test only measured one cognitive ability and its presentation was too narrow to be a general test. Consequently,

these ideas were extended to other cognitive abilities in [14] by the introduction of other 'factors', and the suggestion of using interactive tasks where "rewards and penalties could be used instead", as in reinforcement learning [13].

Similar ideas followed relating compression and intelligence. Compression tests were proposed as a test for artificial intelligence [30], arguing that "optimal text compression is a harder problem than artificial intelligence as defined by Turing's". Nonetheless, the fact that there is a connection between compression and intelligence does not mean that intelligence can be just defined as compression ability (see, e.g., [9] for a full discussion on this).

Later, [27] would propose a notion which they referred to as a "universal intelligence measure" —universal because of its proposed use of a universal distribution for the weighting over environments. The innovation was mainly their use of a reinforcement learning setting, which implicitly accounted for the abilities not only of learning and prediction, but also of planning. An interesting point for making this proposal popular was its conceptual simplicity: intelligence was just seen as average performance in a range of environments, where the environments were just selected by a universal distribution.

While innovative, the universal intelligence *measure* [27] showed several shortcomings stopping it from being a viable *test*. Some of the problems are that it requires a summation over infinitely many environments, it requires a summation over infinite time within each environment, Kolmogorov complexity is typically not computable, disproportionate weight is put on simple environments (e.g., with $1-2^{-7} > 99\%$ of weight put on environments of size less than 8, as also pointed out by [21]), it is (static and) not adaptive, it does not account for time or agent speed, etc

Hernandez-Orallo and Dowe [17] re-visited this to give an intelligence *test* that does not have these abovementioned shortcomings. This was presented as an anytime universal intelligence test. The term *universal* here was used to designate that the test could be applied to any kind of subject: machine, human, non-human animal or a community of these. The term *anytime* was used to indicate that the test could evaluate any agent speed, it would adapt to the intelligence of the examinee, and that it could be interrupted at any time to give an intelligence score estimate. The longer the test runs, the more reliable the estimate (the average reward [16]).

Preliminary tests have since been done [23, 24, 28] for comparing human agents with non-human AI agents. These tests seem to succeed in bringing theory to practice quite seamlessly and are useful to compare the abilities of systems of the same kind. However, there are some problems when comparing systems of different kind, such as human and AI algorithms, because the huge difference of both (with current state-of-the-art technology) is not clearly appreciated. One explanation for this is that (human) intelligence is the result of the adaptation to environments where the probability of other agents (of lower or similar intelligence) being around is very high. However, the probability of having another agent of even a small degree of intelligence just by the use of a universal distribution is discouragingly remote. Even in environments where other agents are included on purpose [15], it is not clear that these agents properly represent a rich 'social' environment. In [18], the so-called Darwin-Wallace distribution is introduced where environments are generated using a universal distribution for multi-agent environments, and where a number of agents that populate the environment are also generated by a universal distribution. The probability of having interesting environments and agents is very low on this first 'generation'. However, if an intelligence test is administered to this population and only those with a certain level are preserved, we may get a second population whose agents will have a slightly higher degree of intelligence. Iterating this process we have different levels for the Darwin-Wallace distribution, where evolution is solely driven (boosted) by a fitness function which is just measured by intelligence tests.

3 THE BASE CASE: THE TURING TEST FOR TURING MACHINES

A recursive approach can raise the odds for environments and tasks of having a behaviour which is attributed to more intelligent agents. This idea of recursive populations can be linked to the notion of *recursive Turing Test* [32, sec. 5.1], where the agents which have succeeded at lower levels could be used to be compared at higher levels. However, there are many interpretations of this informal notion of a recursive Turing Test. The fundamental idea is to eliminate the human reference from the test using recursion —either as the subject that has to be imitated or the judge which is used to tell between the subjects.

Before giving some (more precise) interpretations of a recursive version of the Turing Test, we need to start with the *base case*, as follows (we use TM and UTM for Turing Machine and Universal Turing Machine respectively):

Definition 1 *The imitation game for Turing machines*[4] *is defined as a tuple* $\langle D, B, C, I\rangle$

- *The reference subject* A *is randomly taken as a TM using a distribution* D*.*
- *Subject* B *(the evaluee) tries to emulate* A*.*
- *The similarity between* A *and* B *is 'judged' by a criterion or judge* C *through some kind of* interaction *protocol* I*. The test returns this similarity.*

An instance of the previous schema requires us to determine the distribution D and the similarity criterion C and, most especially, how the interaction I goes. In the classical Turing Test, we know that D is the human population, C is given by a human judge, and the interaction is an open teletype conversation[5]. Of course, other distributions for D could lead to other tests, such as, e.g., a canine test, taking D as a dog population, and judges as other dogs which have to tell which is the member of the species or perhaps even how intelligent it is (for whatever purpose —e.g., mating or idle curiosity).

More interestingly, one possible instance for Turing machines could go as follows. We can just take D as a universal distribution over a reference UTM U, so $p(A) = 2^{-K_U(A)}$, where $K_U(A)$ is the prefix-free Kolmogorov complexity of A relative to U. This means that simple reference subjects have higher probability than complex subjects. Interaction can go as follows. The 'interview' consists of questions as random finite binary strings using a universal distribution $s_1, s_2, ...$ over another reference UTM, V. The test starts by subjects A and B receiving string s_1 and giving two sequences a_1 and b_1 as respective answers. Agent B will also receive what A has output

[4] The use of Turing machines for the reference subject is relevant and not just a way to link two things by their name, Turing. Turing machines are required because we need to define formal distributions on them, and this cannot be done (at least theoretically) for humans, or animals or 'agents'.

[5] This free teletype conversation may be problematic in many ways. Typically, the judge C wishes to steer the conversation in directions which will enable her to get (near-)maximal (expected) information (before the time-limit deadline of the test) about whether or not the evaluee subject B is or is not from D. One tactic for a subject which is not from D (and not a good imitator either) is to distract the judge C and steer the conversation in directions which will give judge C (near-) minimal (expected) information.

immediately after this. Judge C is just a very simple function which compares whether a_1 and b_1 are equal. After one interation, the system issues string s_2. After several iterations, the score (similarity) given to B is calculated as an aggregation of the times a_i and b_i have been equal.

This can be seen as formalisation of the Turing Test where it is a Turing machine that needs to be imitated, and the criterion for imitation is the similarity between the answers given by A and B to the same questions. If subject B cannot be told or instructed about the goal of the test (imitating A) then we can use rewards after each step, possibly concealing A's outputs from B as well.

This test might seem ridiculous at first sight. Some might argue that being able to imitate a randomly-chosen TM is not related to intelligence. However, two issues are important here. First, agent B does not know who A is in advance. Second, agent B tries to imitate A solely from its behaviour.

This makes the previous version of the test very similar to the most abstract setting used for analysing what learning is, how much complexity it has and whether it can be solved. First, this is tantamount to Gold's language identification in the limit [11]. If subject B is able to identify A at some point, then it will start to score perfectly from that moment. While Gold was interested in whether this could be done in general and for every possible A, here we are interested in how well B does this on average for a randomly-chosen A from a distribution. In fact, many simple TMs can be identified quite easily, such as those simple TMs which output the same string independently of the input. Second, and following this averaging approach, Solomonoff's setting is also very similar to this. Solomonoff proved that B could get the best estimations for A if B used a mixture of all consistent models inversely weighted by 2 to the power of their Kolmogorov complexity. While this may give the best theoretical approach for prediction and perhaps for "imitation", it does not properly "identify" A. Identification can only be properly claimed if we have one single model of A which is exactly as A. This distinction between one vs. multiple models is explicit in the MML principle, which usually considers just one single model, the one with the shortest two-part message encoding of said model followed by the data given this model.

There is already an intelligence test which corresponds to the previous instance of definition 1, the C-test, mentioned above. The C-test measures how well an agent B is able to identify the pattern behind a series of sequences (each sequence is generated by a different program, i.e., a different Turing machine). The C-test does not use a query-answer setting, but the principles are the same.

We can develop a slight modification of definition 1 by considering that subject A also tries to imitate B. This might lead to easy convergence in many cases (for relatively intelligent A and B) and would not be very useful for comparing A and B effectively. A significant step forward is when we consider that the goal of A is to make outputs that cannot be imitated by B. While it is clearly different, this is related to some versions of Turing's imitation game, where one of the human subjects pretends to be a machine. While there might be some variants here to explore, if we restrict the size of the strings used for questions and answers to 1 (this makes agreeing and disagreeing equally likely), this is tantamount to the game known as 'matching pennies' (a binary version of rock-paper-scissors where the first player has to match the head or tail of the second player, and the second player has to disagree on the head or tail of the first). Interestingly, this game has also been proposed as an intelligence test in the form of Adversarial Sequence Prediction [20][22] and is related to the "elusive model paradox" [3, footnote 211][4, p 455][5, sec. 7.5].

This instance makes it more explicit that the distribution D over the agents that the evaluee has to imitate or compete with is crucial. In the case of imitation, however, there might be non-intelligent Turing machines which are more difficult to imitate/identify than many intelligent Turing machines, and this difficulty seems to be related to the Kolmogorov complexity of the Turing machine. And linking difficulty to Kolmogorov complexity is what the C-test does. But biological intelligence is frequently biased to social environments, or at least to environments where other agents can be around eventually. In fact, societies are usually built on common sense and common understanding, but in humans this might be an evolutionarily-acquired ability to imitate other humans, but not other intelligent beings in general. Some neurobiological structures, such as *mirror neurons* have been found in primates and other species, which may be responsible of understanding what other people do and will do, and for learning new skills by imitation. Nonetheless, we must say that human unpredictability is frequently impressive, and its relation to intelligence is far from being understood. Interestingly, some of the first analyses on this issue [34][29] linked the problem with the competitive/adversarial scenario, which is equivalent to the matching pennies problem, where the intelligence of the peer is the most relevant feature (if not the only one) for assessing the difficulty of the game, as happens in most games. In fact, matching pennies is the purest and simplest game, since it reduces the complexity of the 'environment' (rules of the game) to a minimum.

4 RECURSIVE TURING TESTS FOR TURING MACHINES

The previous section has shown that introducing agents (in this case, agent A) in a test setting requires a clear assessment of the distribution which is used for introducing them. A general expression of how to make a Turing Test for Turing machines recursive is as follows:

Definition 2 *The recursive imitation game for Turing machines is defined as a tuple $\langle D, C, I\rangle$ where tests and distributions are obtained as follows:*

1. *Set $D_0 = D$ and $i = 0$.*
2. *For each agent B in a sufficiently large set of TMs*
3. *Apply a sufficiently large set of instances of definition 1 with parameters $\langle D_i, B, C, I\rangle$.*
4. *B's intelligence at degree i is averaged from this sample of imitation tests.*
5. *End for*
6. *Set $i = i + 1$*
7. *Calculate a new distribution D_i where each TM has a probability which is directly related to its intelligence at level $i - 1$.*
8. *Go to 2*

This gives a sequence of D_i.

The previous approach is clearly uncomputable in general, and still intractable even if reasonable samples, heuristics and step limitations are used. A better approach to the problem would be some kind of propagation system, such as Elo's rating system of chess [10], which has already been suggested in some works and competitions in artificial intelligence. A combination of a *soft* universal distribution, where simple agents would have slightly higher probability, and a one-vs-one credit propagation system such as Elo's rating (or any other mechanism which returns maximal expected information with a minimum of pairings), could feasibly aim at having a reasonably

good estimate of the relative abilities of a big population of Turing machines, including some AI algorithms amongst them.

What would this rating mean? If we are using the imitation game, a high rating would show that the agent is able to imitate/identify other agents of lower rating well and that it is a worse imitator/identifier than other agents with higher rating. However, there is no reason to think that the relations are transitive and anti-reflexive; e.g., it might even happen that an agent with very low ranking would be able to imitate an agent with very high ranking better than the other way round.

One apparently good thing about this recursion and rating system is that the start-up distribution can be very important from the point of view of heuristics, but it might be less important for the final result. This is yet another way of escaping from the problems of using a universal distribution for environments or agents, because very simple things take almost all the probability —as per section 2. Using difficulty as in the C-test, making adaptive tests such as the anytime test, setting a minimum complexity value [21] or using hierarchies of environments [22] where "an agent's intelligence is measured as the ordinal of the most difficult set of environments it can pass" are solutions for this. We have just seen another possible solution where evaluees (or similar individuals) can take part in the tests.

5 DISCUSSION

The Turing test, in some of its formulations, is a game where an agent tries to imitate another (or its species or population) which might (or might not) be cheating. If both agents are fair, and we do not consider any previous information about the agents (or their species or populations), then we have an imitation test for Turing machines. If one is cheating, we get closer to the adversarial case we have also seen.

Instead of including agents arbitrarily or assuming that any agent has a level of intelligence a priori, a recursive approach is necessary. This is conceptually possible, as we have seen, although its feasible implementation needs to be carefully considered, possibly in terms of rankings after random 1-vs-1 comparisons.

This view of the (recursive) Turing test in terms of Turing machines has allowed us to connect the Turing test with fundamental issues in computer science and artificial intelligence, such as the problem of learning (as identification), Solomonoff's theory of prediction, the MML principle, game theory, etc. These connections go beyond to other disciplines such as (neuro-)biology, where the role of imitation and adversarial prediction are fundamental, such as predator-prey games, mirror neurons, common coding theory, etc. In addition, this has shown that the line of research with intelligence tests derived from algorithmic information theory and the recent Darwin-Wallace distribution are also closely related to this as well. This (again) links this line of research to the Turing test, where humans have been replaced by Turing machines.

This sets up many avenues for research and discussion. For instance, the idea that the ability of imitating relates to intelligence can be understood in terms of the universality of a Turing machine, i.e. the ability of a Turing machine to emulate another. If a machine can emulate another, it can acquire all the properties of the latter, including intelligence. However, in this paper we have referred to the notion of 'imitation', which is different to the concept of Universal Turing machine, since a UTM is defined as a machine such that there is an input that turns it into any other pre-specified Turing machine. A machine which is able to imitate well is a good learner, which can finally identify any pattern on the input and use it to imitate the source. In fact, a good imitator is, *potentially*, very intelligent, since it can, in theory (and disregarding efficiency issues), act as any other very intelligent being by just observing its behaviour. Turing advocated for learning machines in section 7 of the very same paper [37] where he introduced the Turing Test. Solomonoff taught us what learning machines should look like. We are still struggling to make them work in practice and preparing for assessing them.

ACKNOWLEDGEMENTS

This work was supported by the MEC projects EXPLORA-INGENIO TIN 2009-06078-E, CONSOLIDER-INGENIO 26706 and TIN 2010-21062-C02-02, and GVA project PROMETEO/2008/051. Javier Insa-Cabrera was sponsored by Spanish MEC-FPU grant AP2010-4389.

REFERENCES

[1] G. J. Chaitin, 'On the length of programs for computing finite sequences', *Journal of the Association for Computing Machinery*, **13**, 547–569, (1966).

[2] G. J. Chaitin, 'Godel's theorem and information', *International Journal of Theoretical Physics*, **21**(12), 941–954, (1982).

[3] D. L. Dowe, 'Foreword re C. S. Wallace', *Computer Journal*, **51**(5), 523 – 560, (September 2008). Christopher Stewart WALLACE (1933-2004) memorial special issue.

[4] D. L. Dowe, 'Minimum Message Length and statistically consistent invariant (objective?) Bayesian probabilistic inference - from (medical) "evidence"', *Social Epistemology*, **22**(4), 433 – 460, (October - December 2008).

[5] D. L. Dowe, 'MML, hybrid Bayesian network graphical models, statistical consistency, invariance and uniqueness', in *Handbook of the Philosophy of Science - Volume 7: Philosophy of Statistics*, ed., P. S. Bandyopadhyay and M. R. Forster, pp. 901–982. Elsevier, (2011).

[6] D. L. Dowe and A. R. Hajek, 'A non-behavioural, computational extension to the Turing Test', in *Intl. Conf. on Computational Intelligence & multimedia applications (ICCIMA'98), Gippsland, Australia*, pp. 101–106, (February 1998).

[7] D. L. Dowe and A. R. Hajek, 'A computational extension to the Turing Test', *in Proceedings of the 4th Conference of the Australasian Cognitive Science Society, University of Newcastle, NSW, Australia*, (September 1997).

[8] D. L. Dowe and J. Hernandez-Orallo, 'IQ tests are not for machines, yet', *Intelligence*, **40**(2), 77–81, (2012).

[9] D. L. Dowe, J. Hernández-Orallo, and P. K. Das, 'Compression and intelligence: social environments and communication', in *Artificial General Intelligence*, eds., J. Schmidhuber, K.R. Thórisson, and M. Looks, volume 6830, pp. 204–211. LNAI series, Springer, (2011).

[10] A.E. Elo, *The rating of chessplayers, past and present*, volume 3, Batsford London, 1978.

[11] E.M. Gold, 'Language identification in the limit', *Information and control*, **10**(5), 447–474, (1967).

[12] J. Hernández-Orallo, 'Beyond the Turing Test', *J. Logic, Language & Information*, **9**(4), 447–466, (2000).

[13] J. Hernández-Orallo, 'Constructive reinforcement learning', *International Journal of Intelligent Systems*, **15**(3), 241–264, (2000).

[14] J. Hernández-Orallo, 'On the computational measurement of intelligence factors', in *Performance metrics for intelligent systems workshop*, ed., A. Meystel, pp. 1–8. National Institute of Standards and Technology, Gaithersburg, MD, U.S.A., (2000).

[15] J. Hernández-Orallo, 'A (hopefully) non-biased universal environment class for measuring intelligence of biological and artificial systems', in *Artificial General Intelligence, 3rd Intl Conf*, ed., M. Hutter et al., pp. 182–183. Atlantis Press, Extended report at http://users.dsic.upv.es/proy/anynt/unbiased.pdf, (2010).

[16] J. Hernández-Orallo, 'On evaluating agent performance in a fixed period of time', in *Artificial General Intelligence, 3rd Intl Conf*, ed., M. Hutter et al., pp. 25–30. Atlantis Press, (2010).

[17] J. Hernández-Orallo and D. L. Dowe, 'Measuring universal intelligence: Towards an anytime intelligence test', *Artificial Intelligence Journal*, **174**, 1508–1539, (2010).

[18] J. Hernández-Orallo, D. L. Dowe, S. España-Cubillo, M. V. Hernández-Lloreda, and J. Insa-Cabrera, 'On more realistic environment distributions for defining, evaluating and developing intelligence', in *Artificial General Intelligence*, eds., J. Schmidhuber, K.R. Thórisson, and M. Looks, volume 6830, pp. 82–91. LNAI, Springer, (2011).
[19] J. Hernández-Orallo and N. Minaya-Collado, 'A formal definition of intelligence based on an intensional variant of Kolmogorov complexity', in *Proc. Intl Symposium of Engineering of Intelligent Systems (EIS'98)*, pp. 146–163. ICSC Press, (1998).
[20] B. Hibbard, 'Adversarial sequence prediction', in *Proceeding of the 2008 conference on Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, pp. 399–403. IOS Press, (2008).
[21] B. Hibbard, 'Bias and no free lunch in formal measures of intelligence', *Journal of Artificial General Intelligence*, **1**(1), 54–61, (2009).
[22] B. Hibbard, 'Measuring agent intelligence via hierarchies of environments', *Artificial General Intelligence*, 303–308, (2011).
[23] J. Insa-Cabrera, D. L. Dowe, S. España-Cubillo, M. Victoria Hernández-Lloreda, and José Hernández-Orallo, 'Comparing humans and ai agents', in *AGI: 4th Conference on Artificial General Intelligence - Lecture Notes in Artificial Intelligence (LNAI)*, volume 6830, pp. 122–132. Springer, (2011).
[24] J. Insa-Cabrera, D. L. Dowe, and José Hernández-Orallo, 'Evaluating a reinforcement learning algorithm with a general intelligence test', in *CAEPIA - Lecture Notes in Artificial Intelligence (LNAI)*, volume 7023, pp. 1–11. Springer, (2011).
[25] A. N. Kolmogorov, 'Three approaches to the quantitative definition of information', *Problems of Information Transmission*, **1**, 4–7, (1965).
[26] S. Legg and M. Hutter, 'Tests of machine intelligence', in *50 years of artificial intelligence*, pp. 232–242. Springer-Verlag, (2007).
[27] S. Legg and M. Hutter, 'Universal intelligence: A definition of machine intelligence', *Minds and Machines*, **17**(4), 391–444, (November 2007).
[28] S. Legg and J. Veness, 'An Approximation of the Universal Intelligence Measure', in *Proceedings of Solomonoff 85th memorial conference*. Springer, (2012).
[29] D. K. Lewis and J. Shelby-Richardson, 'Scriven on human unpredictability', *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, **17**(5), 69 – 74, (October 1966).
[30] M. V. Mahoney, 'Text compression as a test for artificial intelligence', in *Proceedings of the National Conference on Artificial Intelligence, AAAI*, pp. 970–970, (1999).
[31] G. Oppy and D. L. Dowe, 'The Turing Test', in *Stanford Encyclopedia of Philosophy*, ed., Edward N. Zalta. Stanford University, (2011). http://plato.stanford.edu/entries/turing-test/.
[32] P. Sanghi and D. L. Dowe, 'A computer program capable of passing IQ tests', in *4th Intl. Conf. on Cognitive Science (ICCS'03), Sydney*, pp. 570–575, (2003).
[33] A.P. Saygin, I. Cicekli, and V. Akman, 'Turing test: 50 years later', *Minds and Machines*, **10**(4), 463–518, (2000).
[34] M. Scriven, 'An essential unpredictability in human behavior', in *Scientific Psychology: Principles and Approaches*, eds., B. B. Wolman and E. Nagel, 411–425, Basic Books (Perseus Books), (1965).
[35] J. R. Searle, 'Minds, brains and programs', *Behavioural and Brain Sciences*, **3**, 417–457, (1980).
[36] R. J. Solomonoff, 'A formal theory of inductive inference', *Information and Control*, **7**, 1–22, 224–254, (1964).
[37] A. M. Turing, 'Computing machinery and intelligence', *Mind*, **59**, 433–460, (1950).
[38] C. S. Wallace, *Statistical and Inductive Inference by Minimum Message Length*, Information Science and Statistics, Springer Verlag, May 2005. ISBN 0-387-23795X.
[39] C. S. Wallace and D. M. Boulton, 'An information measure for classification', *Computer Journal*, **11**(2), 185–194, (1968).
[40] C. S. Wallace and D. L. Dowe, 'Minimum message length and Kolmogorov complexity', *Computer Journal*, **42**(4), 270–283, (1999).

What language for Turing Test in the age of *qualia*?

Francesco Bianchini[1], **Domenica Bruni**[2]

Abstract. What is the most relevant legacy by Turing for epistemology of Artificial Intelligence (AI) and cognitive science? Of course, we could see it in the ideas set out in his well-known article of 1950, *Computing Machinery and Intelligence*. But how could his imitation game, and its following evolution in what we know as Turing Test, still be so relevant? What we want to argue is that the nature of imitation game as a method for evaluating research on intelligent artifacts, has not its core specifically in (natural) language capability as a way of showing the presence of intelligence in a certain entity, but in the interaction between human being and machines. Human-computer interaction is a particular field in information science for many important practical respects, but interaction between human being and machines is the deepest sense of Turing's ideas on evaluation of intelligent behavior and entities, within and beyond its connection with natural language. And from this point of view it could be methodologically and epistemologically useful for further research in every discipline involving machine and artificial artifacts, especially as concerns the very current subject of consciousness and *qualia*. In what follows we will try to argue such a perspective by showing some field in which interaction, in connection with different sorts of language, could be of interest in the spirit of Turing's 1950 article.

1 TURING, LANGUAGE AND INTERACTION

One of the most interesting idea by Turing was a based-on-language test for proving the intelligence, or the intelligent behavior, of a program. In Turing's terms, it is a machine showing an autonomous and self-produced intelligent behavior. Actually, Turing never spoke about a test, but just about an imitation game, using the concept of imitation as an intuitive concept. This is a typical way of thinking as regards Turing, though, who had provided a method for catching the notion of computable function in a mechanical way through a set of intuitive concepts about fifteen years before [24]. Likewise the case of computation theory, the Turing's aim in 1950 article was to deal with a very notable subject in the easiest and most straightforward manner, and avoiding the involvement with more complex and specific theoretical structures based on field-dependent notions.

In the case of imitation game the combination of the notion of "imitation" and of the use of natural language allowed Turing to express a paradigmatic method for evaluating artificial products, but gave rise as well to an endless debate all over the last sixty years about the suitableness of this kind of testing artificial *intelligence*. Leaving aside the problem concerning the correct interpretation of the notion of "imitation", we may ask first whether the role of language in the test is fundamental or it is just connected to the spirit of the period in which Turing wrote his paper, that is within the current behaviorist paradigm in psychology and in the light of the natural language centrality in the philosophy of twentieth century. In other terms, why did Turing choose natural language in order to build a general frame for evaluating the intelligence of artificial, programmed artifacts? Is such a way of thinking (and researching) still useful? And, if so, what can we say about it in relation with further research in this field?

As we said, the choice of natural language had the purpose to put the matter in an intuitive manner. We human beings usually ascribe intelligence to other human beings through linguistic conversations, mostly carrying out in a question-answer form. Besides, Turing himself asserts in 1950 article that such a method «has the advantage of drawing a fairly sharp line between the physical and the intellectual capacities of a man» [26]. This is the ordinary explanation of Turing's choice. But it is also true that, in a certain sense, the very first enunciation of the imitation game is in another previous work by Turing, where, ending his exposition on machine intelligence, he speaks about a «little experiment» regarding the possibility of a chess game between two human beings (A and C), and between a human being (A) and a paper machine worked by a human being (B). Turing asserts that if «two rooms are used with some arrangement for communicating moves, and a game is played between C and either A or the paper machine […] C may find it quite difficult to tell which he is playing. (This is a rather idealized form of an experiment I have actually done.)» [25].

Such a brief sketch of the imitation game in 1948 paper is not surprising because that paper is a sort of first draft of the Turing's ideas of 1950 paper, and it is even more considerable for some remarks, for example, on self-organizing machines or on the possibility of machine learning. Moreover, it is not surprising that Turing speaks about machines referring to them as paper machines, namely just for their logical, abstract structure. It is another main Turing's theme, that remembers the human *computor* of 1936 paper. What is interesting is the fact that the first, short outline of imitation game is not based on language, but on a subject that is more early-artificial-intelligence-like, that is, chess game. So, (natural) language is not necessary for imitation game from the point of view of Turing, and yet the ordinary explanation of Turing's choice for language is still valid within such a framework. In other terms, Turing was aware not only that there are other domains in which a machine can apply itself autonomously – a trivial fact – but also that such domains are as enough good as natural language for imitation game. Nevertheless, he choose natural language as paradigmatic.

What conclusions can we draw from such remarks? Probably two ones. First, Turing was pretty definitely aware that the evaluation of artificial intelligence (AI) products, in a broad

[1] Dept. of Philosophy, University of Bologna. Email: francesco.bianchini5@unibo.it

[2] Dept. of Cognitive Science, University of Messina, Email: dbruni@unime.it

sense, would be a very difficult subject, maybe the more fundamental as regards the epistemology of AI and cognitive science, even if, obviously, he didn't use such terms in 1950. Secondly, that the choice of language and the role of language in imitation game are even more subtle than the popular culture and the AI tradition usually assert. As a matter of fact, he did not speak about natural language in general but of a "question-answer method", a method that involves *communication*, not just language processing or producing. So, from this point of view it seems that, for Turing, natural language processing or producing are just some peculiar human cognitive abilities among many other ones, and are not basic for testing intelligence. What is basic for such a task is communication or, to use another, more inclusive term, *interaction*. But a specification is needed. We are not maintaining that the capability of using language is not a cognitive feature, but that in Turing's view interaction is the best way in order to detect intelligence, and *language* interaction, by means of question-answer method, is perhaps the most intuitive form of interaction for human beings. No interaction is tantamount to no possibility to identify intelligence, and for such a purpose one of the two poles of interaction must be a human being[3].

Furthermore, the «question and answer method seems to be suitable for introducing almost anyone of the fields of human endeavour that we wish to include» [26] and, leaving aside the above-mentioned point concerning the explicit Turing's request to penalize in no way machines or human beings for their unshared features, we could consider it as the main aim of Turing, namely generalizing the intelligence testing. Of course, such an aim anticipates one of the mainstream of the following rising AI[4], but it has an even wider range. Turing was not speaking, indeed, about problem solving, but trying to formulate a criterion and a method to show and identify machine intelligent behavior in different-field interaction with human beings. So, language communication seems to become *both* a lowest common denominator for every field in which it is possible testing intelligence *and*, at the same time, a way to cut single field or domain for testing intelligence from the point of view of interaction. Now we will consider a few of them, in order to investigate and discuss whether they could be relevant for *qualia* problem.

[3] A similar way of thinking seems to be suggested, as regards specifically natural language, by an old mental experiment formulated by Putnam, in which he imagines a human being learning by heart a passage in a language he did not know and then repeating it in a sort of stream of consciousness. If a telepath, knowing that particular language, could perceive the stream of consciousness of the human being who has memorized the passage, the telepath could think the human being knows that language, even though it is not so. What does it lack in the scene described in the mental experiment? A real interaction. As a matter of fact, the conclusion of Putnam himself is that: «the understanding, then, does not reside in the words themselves, nor even in the appropriateness of the whole sequence of words and sentences. It lies, rather, in the fact that an *understanding* speaker can *do things* with the words and sentences he utters (or thinks in his head) *besides* just utter them. He can answer questions, for example […].» [19]. And it appears to be very close to what Turing thought more than twenty years before.

[4] For example, consider the target to build a General Problem Solver pursued by Newell, Shaw and Simon for long [15, 16].

2 LANGUAGE TRANSLATION AS CULTURAL INTERACTION

A first field in which language and interaction are involved is language translation. We know that machine translation is a very difficult target of computer science and AI since their origins up to nowadays. The reason is that translation usually concerns two different natural languages, two tongues, and it is not a merely act of substitution. On the contrary, translation involves many different levels of language: syntactic and semantic levels, but also cultural and stylistic levels, that are very context-dependent. It is very difficult for a machine to find the correct word or expression to yield in a specific language what is said in another language. Many different approaches in this field, especially from computational linguistic, are available to solve the problem of a good translation. But anyway, it is an operation that still remains improvable. As a matter of fact, if we consider some machine translation tools like Google Translator, there are generally syntactic and semantic problems in every product of such tools, even if, maybe, the latter are larger than the former. So, how can we test intelligence in this field concerning language? Or, in other terms, what could be a real test for detecting intelligence as regards translation? A tool improvement could be not satisfying. We could think indeed that, with the improvement of machine translation tools, we could have better and better outcomes in this field, but what we want is not a collection of excellent texts, from the point of view of translation. What we want is a sort of justification of the word choice in the empirical activity of translation. If we could have a program that is able to justify its choosing of words and expressions in the act of translation, we could consider that the problem of a random good choice of a word or of an expression is evaded.

In a dialogue, a personal tribute to Alan Turing, Douglas Hofstadter underlines a similar view. Inspired by the two little snippets of Turing's 1950 article [26], Hofstadter builds a (fictitious) conversation between a human being and a machine in order to show the falsity of simplistic interpretations of Turing Test, that he summarizes in the following way: «*even if some AI program passed the full Turing Test, it might still be nothing but a patchwork of simple-minded tricks, as lacking in understanding or semantics as is a cash register or an automobile transmission*» [10]. In his dialogue, Hofstadter tries to expand the flavor of the second Turing snippet, where Mr Pickwick is compared to a winter's day [26]. The conversation by Hofstadter has translation as the main topic, in particular poetry translation. Hofstadter wants to show how complex such a subject is and that it is very difficult that a program could have a conversation of that type with a human being, and thus pass the Turing Test. By reversing perspective, we can consider translation one of the language field in which, in the future, it could be fruitful testing machine intelligence. But we are not merely referring to machine translation. We want to suggest the a conversation on a translation subject could be a target for a machine. Translation by itself, indeed, concerns many cultural aspects, as we said before, and the understanding and justification of what term or expression is suitable in a specific context of a specific language could be a very interesting challenge for a program, that would imply the knowledge of the cultural context of a specific language by the program, and

therefore the implementation of mechanisms for representing and handling two different language contexts.

In Hofstadter's dialogue, much attention is devoted to the problem from a poetic point of view. We can have a flavour of the general issues involved by considering an extract from the dialogue, which is between two entities, a Dull Rigid Human and an Ace Mechanical Translator:

«*DRH*: Well, of course, being an advanced AI program, you engaged in a highly optimized heuristic search.

AMT: For want of a better term, I suppose you could put it that way. The constraints I found myself under in my search were, of course, both semantic and phonetic. Semantically, the problem was to find some phrase whose evoked imagery was sufficiently close to, or at least reminiscent of, the imagery evoked by *croupir dans ton lit*. Phonetically, the problem was a little trickier to explain. Since the line just above ended with "stir", I needed an "ur" sound at the end of line 6. But I didn't want to abandon the idea of hyphenating right at that point. This meant that I needed two lines that matched this template:

Instead of …ur…ing ……bed

where the first two ellipses stand for consonants (or consonant clusters), and the third one for "in" or "in your" or something of the sort. Thus, I was seeking gerunds like "lurking", "working", "hurting", "flirting", "curbing", "squirming", "bursting", and so on — actually, a rather rich space of phonetic possibilities.

DRH: Surely you must have, within your vast data bases, a thorough and accurate hyphenation routine, and so you must have known that the hyphenations you propose — "lur-king", "squir-ming", "bur-sting", and so forth — are all illegal…

AMT: I wish you would not refer to my knowledge as "your vast data bases". I mean, why should that quaint, old-fashioned term apply to *me* any more than to *you*? But leaving that quibble aside, yes, of course, I knew that, strictly speaking, such hyphenations violate the official syllable boundaries in the eyes of rigid language mavens like that old fogey William Safire. But I said to myself, "Hey, if you're going to be so sassy as to hyphenate a word across a line-break, then why not go whole hog and hyphenate in a sassy spot *inside* the word?"» [10].

Poetry involves metrical structures, rhymes, assonances, alliterations and many other figures of speech [10]. But, they constitute some constraints that are easily mechanizable, by means of the appropriate set of data bases. In fact, a machine could be faster than a human being in finding, for example, every word rhyming with a given one. So the problem is not if we have to consider poetry or prose translation, and their differences, but that of catching the cultural and personal flavor of the text's author, within a figure of speech scheme or not. Poetry just has some further, but mechanizable, constraints. So, what remains outside such constraints? Is it the traditional idea of an intentionality of terms? We do not think that things are those. The notion of intentionality seems always to involve a first-person, subjective point of view that is undetectable in a machine, as a long debate of last thirty years seems to show. But if we consider the natural development of intentionality problem, that of *qualia*, (as subjective conscious experiences that we are able to express with words), maybe we could have a better problem and find a better field of investigation in considering translation as a sort of *qualia* communication. In other terms, a good terminological choice and a good justification of such a choice could be a suitable method for testing intelligence, even in its capability to express and understand *qualia*. And this could be a consequence of the fact that, generally speaking, translation is a sort of communication, a communication of contents from a particular language to another particular language; and in the end a context interaction.

3 INTERACTION BETWEEN MODEL AND REALITY

Another field in which the notion of interaction could be relevant from the point of view of the Turing Test is that of scientific discovery. In the long development of machine learning some researchers implemented programs that are able to carry out generalizations from data structures within a specific scientific domain, namely scientific laws[5]. Even thought they are very specific laws, they are (scientific) laws in all respects. Such programs were based on logic method and, indeed, they could only arrive to a generalization from data structures and they were not able to obtain their outcomes from experimental conditions. More recently, other artificial artifacts have been built in order to fill such a gap. For example, ADAM [8] is a robot programmed for carrying out outcomes in genetics with the possibility of autonomously managing real experiments. It has a logic-based knowledge base that is a model of metabolism, but it is able as well to plan and run experiments to confirm or disconfirm some hypotheses within a research task. In particular, it could set up experimental conditions and situations with a high level of resource optimization for investigating gene expression and associating one or more genes to one protein. The outcome is a (very specific but real) scientific law, or a set of them. We could say that ADAM is a theoretical and practical machine. It formulates a number of hypotheses of gene expression using its knowledge bases, that includes all that we already know about gene expression from a biological point of view. It does the experiments to confirm or disconfirm every hypothesis, and then it carries out a statistical analysis for evaluating the results. So, is ADAM a perfect scientist, an autonomous intelligent artifacts in the domain of science?

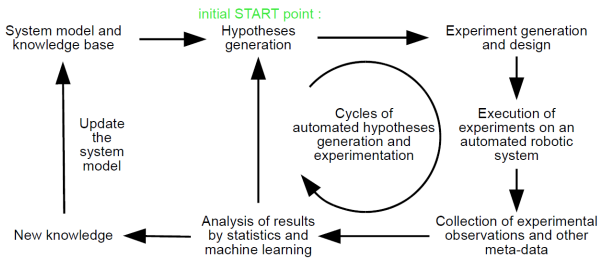


Figure 1. Diagram of the hypotheses generation–experimentation cycle for the production of new scientific knowledge, on which ADAM is based (from [21]).

[5] For example GOLEM. For some outcomes of it, see [14]; for a discussion see [5].

Of course, it is true that its outcomes are original in some cases; and it is also true that its creators, its programmers do not see in it a substitute for scientists, but only an assistant for human scientists, even though a very efficient one, at least at the current phase of research, likewise it happens in other fields like chess playing and music. What does lack ADAM to become a scientist in? We could say that it lacks in the possibility of controlling or verifying its outcomes from different points of view, for example from an interdisciplinary perspective. But it seems a mere practical limit, surmountable with a lot of additional scientific knowledge of different domains, given that it has the concrete possibility to do experiments. Yet, as regards such a specific aspect, what is the reach of ADAM – or other programs devoted to scientific discovery, like EVE, specialized in pharmaceutical field – in conducting experiments? Or, that is the same thing, how far could it get in formulating hypotheses? It all seems to depend on its capacity of interaction with the real world. And so we could say that in order to answer the question if ADAM or other similar artificial artifacts are intelligent, we have to consider not only the originality of their outcomes, but also their creativity in the hypothesis formulation, task that is strictly dependent on its practical interaction with the real world. Is this a violation of what Turing said we have not to consider in order to establish if a machine is intelligent, namely its "physical" difference from human beings? We think not. We think that interaction between a model of reality and reality itself from a scientific point of view is the most important aspect in scientific discovery and it could be in the future one of the way in which evaluate the results of artificial artifacts and their intelligence. As a matter of fact, science and scientific discovery take place in a domain in which knowledge and methods are widely structured and the invention of new hypotheses and theories could reveal itself as a task of combination of previous knowledge, even expressed in some symbolic language, more than a creation from nothing. And the capability to operate such a combination could be the subjective perspective, the first person point of view of future machines.

4 EMOTION INTERACTING: THE CASE OF LOVE

Another field in which the notion of interaction could be relevant from the point of view of Turing Test are emotions, their role in the interaction with the environment and the language to transmit the emotions. Emotions are cognitive phenomena. It is not possible to characterize them as irrational dispositions, but they provide with all the necessary information about the word around us. The emotions are a way to relate the environment and other individuals. Emotions are probably a necessary condition for our mental life [2, 6]. They show us our radical dependence on the natural and social environment.

One of the most significant cognitive emotions is love. Since antiquity, philosophers have considered love as a crucial issue in their studies. Modern day psychologists have discussed its dynamics and dysfunctions. However, it has rarely been investigated as a genuine human cognitive phenomenon. In its most common sense, love has been considered in poetry, philosophy, and literature, as being something universal, but at the same time, as a radically subjective feeling. This ambiguity is the reason why love is such a complicated subject matter. Now, we want to argue that love, by means of its rational character, can be studied in a scientific way. According to the philosophical tradition, human beings are rational animals. However, the same rationality guides us in many circumstances, sometimes creates difficult puzzles. Feelings and emotions, like love, fortunately are able to offer an efficient reason for action.

Even if what "love" is defies definition, it remains a crucial experience in the ordinary life of human beings. It participates in the construction of human nature and in the construction of an individual's identity. This is shown by the universality of the feeling of love across cultures. It is rather complicated to offer a precise definition of "love", because its features include emotional states, such as tenderness, commitment, passion, desire, jealousy, and sexuality. Love modifies people's way of thinking and acting, and it is characterized by a series of physical symptoms. In fact, love has often been considered as a type of mental illness. How many kinds of love are there? In what relation are they?

Over the past decades many classifications of love have been proposed. Social psychologists such as Berscheid and Walster [1], for example, in their cognitive theory of emotion, propose two stages of love. The former has to do with a state of physiological arousal and it is caused by the presence of positive emotions, like sexual arousal, satisfaction, and gratification, or by negative emotions, such as fear, frustration, or being rejected. The second stage of love is called "tagging", i.e., the person defines this particular physiological arousal as a "passion" or "love". A different approach is taken by Lee [12] and Hendrick [7, 9]. Their interest is to identify the many ways we have for classifying or declining love. They focus their attention on love styles, identifying six of them: *Eros, Ludus, Mania, Pragma*, *Storge* and *Agape*. *Eros* (passionate love) is the passionate love which gives central importance to the sexual and physical appearance of the partner; *Ludus* (game-playing love) is a type of love exercised as a game that does not lead to a stable, lasting relationship; *Mania* (possessive, dependent love) is a very emotional type of love which is identified with the stereotype of romantic love; *Pragma* (logical love) concerns the fact that lovers have a concrete and pragmatic sense of the relationship, using romance to satisfy their particular needs and dictating the terms of them; *Storge* (friendship-based love) is a style in which the feeling of love toward each other grows very slowly. Finally, it is possible to speak of *Agape* (all-giving selfless love) characterized by a selfless, spiritual and generous love, something rarely experienced in the lifetime of individuals. Robert Sternberg [20] offers a graphical representation of love called the "triangle theory". The name stems from the fact that the identified components are the vertices of a triangle. The work of the Yale psychologist deviates from previous taxonomies, or in other words, from the previous attempts made to offer a catalogue of types of existing love. The psychological elements identified by Sternberg to decline feelings of love are three: intimacy, passion, decision/commitment. The different forms of love that you may encounter in everyday life would result from a combination of each of these elements or the lack of them. Again, in the study and analysis of the feeling of love we encounter a list of types of love: non-love, affection, infatuation, empty love, romantic love, friendship, love, fatuous love, love lived.

Philosophers, fleeing from any kind of taxonomy, approach the feeling of love cautiously, surveying it and perhaps even fearing it. Love seems to have something in common with the

deepest of mysteries, i.e. the end of life. It leads us to question, as death does, the reality around us as well as ourselves, in the hope that something precious and important not pass us by. But love is also the guardian of an evil secret that is revealed, which consists in the nonexistence of the love object, in that it is nothing but a projection of our own desires. Love is, according to Arthur Schopenhauer, a sequence of actions performed by those who know perfectly that there is a betrayal in that it does nothing else but carry out the painful event which life consists in. Thus, love, too, has its Maya veil, and once torn down, what remains? What remains is the imperative of the sexual reproduction of the species instinct.

Human nature has for Harry G. Frankfurt [4] two fundamental characteristics: rationality and the capacity to love. Reason and love are the regulatory authorities that guide the choices to be made, providing the motivation to do what we do and constraining it by creating a space which circumscribes or outlines the area in which we can act. On one hand, the ability to reflect and think about ourselves leads to a sort of paralysis. The ability to reflect, indeed, offers the tools to achieve our desires, but at the same time, is often an impediment to their satisfaction, leading to an inner split. On the other, the ability to love unites all our fragments, structuring and directing them towards a definite end. Love, therefore, seems to be involved in integration processes of personal identity.

In *The Origin of species* [3] Charles Darwin assigned great importance to sexual selection, arguing that language, in its gradual development, was the subject of sexual selection, recognizing in it features of an adaptation that we could call unusual (such as intelligence or morality). The dispute that has followed concerning language and its origins has ignited the minds of many scholars and fueled the debate about whether language is innate or is, on the contrary, a product of learning. Noam Chomsky has vigorously fought this battle against the tenets of social science supporting that language depends on an innate genetic ability.

Verbal language is a communication system far more complex than other modes of communication. There are strong referential concepts expressed through language that are capable of building worlds. Similar findings have been the main causes of the perception of language within the community of scholars, as something mysterious, something that appeared suddenly in the course of our history. For a long time arguments concerning the evolution of language were banned and the idea that a similar phenomenon could be investigated and argued according to the processes that drive the evolution of the natural world were considered to be of no help in understanding the complex nature of language. Chomsky was one of the main protagonists of this theoretical trend. According to Chomsky, the complex nature of language is that it can be understood only through a formal and abstract approach such as the paradigm of generative grammar. This theoretical position puts out the possibility of a piecemeal approach to the study of language and the ability to use the theory of evolution to get close to understanding it. Steven Pinker and Paul Bloom, two well-known pupils of Chomsky, in an article entitled "Natural Language and Natural Selection", renewed the debate on the origin of language, stating that it is precisely the theory of evolution that presents the key to explaining the complexity of language. A fascinating hypothesis on language as a biological adaptation is that which considers it an important feature in courtship. Precisely for this reason it would have been subject to sexual selection [13]. A good part of courtship has a verbal nature. Promises, confessions, stories, statements, requests for appointments are all linguistic phenomena. In order to woo, find the right words, find the right tone of voice and the appropriate arguments, you need to employ language.

Even the young mathematician Alan Turing utilized the courtship form to create his imitation game with the aim of finding an answer to a simple – but only in appearance – question ("can machines think?"). Turing formulated and proposed a way to establish it by means of a game that has three protagonists as subject: a man, a woman and an interrogator. The man and woman are together in one room, in another place is the interrogator and communication is allowed through the use of a typewriter. The ultimate goal of the interrogator is to identify if on the other side there is a man or a woman. The interesting part concerns what would happen if in the man's place a computer was put that could simulate the communicative capabilities of a human being. As we mentioned before, the thing that Turing emphasizes in this context is that the only point of contact between human being and machine communication is language. If your computer is capable of expressing a wide range of linguistic behavior appropriate to the specific circumstances it can be considered intelligent. Among the behaviors to be exhibited, Turing insert kindness, the use of appropriate words, and autobiographical information. The importance of transferring to whoever stands in front of us autobiographical information, coating therefore the conversation with a personal and private patina, the expression of shared interests, the use of kindness and humor, are all ingredients typically found in the courtship rituals of human beings. It is significant that a way in which demonstrating the presence of a real human being passed through a linguistic courtship, a mode of expression that reveals the complex nature of language and the presence within it of cognitive abilities. Turing asks: "Can machines think?", and we might answer: "Maybe, if they could get a date on a Saturday evening".

To conclude, in the case of a very particular phenomenon such as love, one of the most intangible emotions, Turing shoves us to consider the role of language as fundamental. But love is a very concrete emotion as well, because of its first person perspective. Nevertheless, in order to communicate it also we human beings are compelled to express it by words in the best way we can, and at the same time we have just language for understanding love emotion in other entities (of course, human beings), together with every real possibility of making mistake and deceiving ourselves. And so, if we admit the reality of this emotion also from a high level cognitive point of view, that involves intelligence and rationality, we have two consequences. The first one is that just interaction reveals love; the second one is that just natural language interaction, made of all the complex concepts that create a bridge between our feelings and the ones of another human being, reveals the *qualia* of the entity involved in a love exchange. Probably that is why Turing wanders through that subject in his imitation game. And probably the understanding of this kind of interaction could be, in the future, a real challenge for artificial artifacts provided with "*qualia* detecting sensor", that cannot be so much different from *qualia* itself.

5 A TURING TEST FOR HUMAN (BEING) BRAIN

A last way in which we could see interaction (connected to language) as relevant for testing intelligence in machines needs two perspective reversals. The first one concerns the use of Turing-Test-like methods to establish the presence of (a certain level of) consciousness in unresponsive brain damage patients. As a matter of fact, such patients are not able to use natural language for communicating as human beings usually do. So researchers try to find signs of communications that are different from languages, like blinks of an eyelid, eye-tracking, simple command following, response to pain, and they try at the same time to understand if they are intentional or automatic [22]. In such cases, neurologists are looking for signs of intelligence, namely of the capability of using intentionally cognitive faculties through a behavioral method that overturns the one of Turing. In the case of machines and Turing Test, natural language faculty is the evidence of the presence of intelligence in machines; in the case of unresponsive brain damage patients, scientists assume that patients were able to communicate through natural language before damage, and so that they were and are intelligent because intelligence is a human trait. Thus, they look for bodily signs to establish a communication that is forbidden through usual means.

This is even more relevant if we consider vegetative state patient, that are not able to perform any body movement. In the last years, some researchers supposed that it is possible to establish a communication with vegetative state patients, a communication that would show also a certain level of consciousness, by means of typical neuroimaging techniques, like fMRI and PET [17][6]. In short, through such experiments they observed that some vegetative state patients, unable to carry out any body response, had a brain activation very similar to that of healthy human beings when they were requested with auditory instructions to imagine themselves walking through one's house or playing tennis. Even though the interpretation of such outcomes is controversial, because of problems regarding neuroimaging methodology and the nature itself of conscious activity, if we accept them, they would prove perhaps the presence of a certain level of consciousness in this kind of patients, namely the presence of consciousness in mental activities. They would prove, thus, the presence of intentionality in the patient response, and not only of cognitive processes or activities, that could be just cognitive "island" of mental functioning [11].

Such experimental outcomes could be very useful for building new techniques and tools of brain-computer interaction for people who are no longer able to communicate by natural language and bodily movements, even though there are many problems that have still to be solved from a theoretical and epistemological point of view as regards the methodology and the interpretations of such results [23]. Is it a real communication? Are those responses a sign of awareness? Could those responses be real answers to external request?

Yet, what is important for our argumentation is the possibility of back-transferring these outcomes to machines, and this is the second reversal we mentioned before. As a matter of fact, these experiments are based on the assumption that also human beings are machines and that communication is interaction between mechanical parts, also in the case of subjective, phenomenal experiences, that are evoked by means of language, but without external signs. So, the challenging question is: is it possible to find a parallel in machines? Is it possible to re-create in artificial artifacts this kind of communication that is not behavioral, but is still mechanical and detectable inside machines – virtual or concrete mechanisms – and is simultaneously a sign of consciousness and awareness in the sense of *qualia*? Is this sort of (non-natural-language) communication, if any, a way in which we could find *qualia* in programs or robots? Is it the sort of interaction that could lead us to the feeling of machines?

[6] For a general presentation and discussion see also [18, 23].

REFERENCES

[1] E. Berscheid, E. Walster, *Interpersonal Attraction*, Addison-Wesley, Boston, Mass., 1978.

[2] A.R. Damasio, *Descartes' Error: Emotion, Reason, and the Human Brain*, Putnam Publishing, New York, 1994.

[3] C. Darwin, *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*, Murray, London, 1859.

[4] H.G. Frankfurt, *The Reasons of Love*, Princeton University Press, Princeton, 2004.

[5] D. Gillies, *Artificial Intelligence and Scientific Method*, Oxford University Press, Oxford, 1996.

[6] P. Griffith, *What emotions really are. The Problem of Psychological Categories*, Chicago University Press, Chicago, 1997.

[7] C. Hendrick, S. Hendrick, 'A Theory and a Method of Love', *Journal of Personality and Social Psychology*, **50**, 392–402, (1986).

[8] R.D.King, J. Rowland, W. Aubrey, M. Liakata, M. Markham, L.N. Soldatova, K.E. Whelan, A. Clare, M. Young, A. Sparkes, S.G. Oliver, P. Pir, 'The Robot Scientist ADAM', *Computer*, **42, 8**, 46–54, (2009).

[9] C. Hendrick, S. Hendrick, *Romantic Love*, Sage, California, 1992.

[10] D.R. Hofstadter, *Le Ton beau de Marot*, Basic Books, New York, 1997.

[11] S. Laureys, 'The neural correlate of (un)awareness: lessons from the vegetative state', *Trends in Cognitive Sciences*, **9, 12**, 556–559, (2005).

[12] J. Lee, *The Colors of Love*, Prentice-Hall, Englewood Cliffs, 1976.

[13] G.F. Miller, *The Mating Mind. How Sexual Choice Shaped the Evolution of Human Nature*, Anchor Books, London, 2001.

[14] S. Muggleton, R.D. King, M.J.E. Sternberg, 'Protein secondary structure prediction using logic-based machine learning', *Protein Engineering*, **5, 7**, 647–657, (1992).

[15] A. Newell, J.C. Shaw, H.A. Simon, 'Report on a general problem-solving program', *Proceedings of the International Conference on Information Processing*, pp. 256–264, (1959).

[16] A. Newell, H.A. Simon, *Human problem solving*, Prentice-Hall, Englewood Cliffs, NJ, 1972.

[17] A.M. Owen, N.D. Schiff, S. Laureys, 'The assessment of conscious awareness in the vegetative state', in S. Laureys, G. Tononi (eds.), *The Neurology of Consciousness*, Elsevier, pp. 163–172, 2009.

[18] A.M. Owen N.D. Schiff, S. Laureys, 'A new era of coma and consciousness science', *Progress in Brain Research*, **177**, 399–411, (2009).

[19] H. Putnam, *Mind, Language and Reality. Philosophical Papers*, Vol. 2. Cambridge University Press, Cambridge, 1975.

[20] R. Sternberg, 'A Triangular Theory of Love', *Psychological Review*, **93**, 119–35, (1986).

[21] A. Sparkes, W. Aubrey, E. Byrne, A. Clare, M.N. Khan, M. Liakata, M. Markham, J. Rowland, L.N. Soldatova, K.E. Whelan, M. Young, R.D. King, 'Towards Robot Scientists for autonomous scientific discovery', *Automated Experimentation*, **2:1**, (2010).

[22] J.F. Stins, 'Establishing consciousness in non-communicative patients: A modern day version of the Turing Test', *Consciousness and Cognition*, **18, 1**, 187–192, (2009).
[23] J.F. Stins, S. Laureys, 'Thought translation, tennis and Turing tests in the vegetative state', *Phenomenology and Cognitive Science*, **8**, 361–370, (2009).
[24] A.M. Turing, 'On Computable Numbers, with an Application to the Entscheidungsproblem', *Proceedings of the London Mathematical Society*, **42**, 230–265, (1936); reprinted in: J. Copeland (ed.), *The essential Turing*, Oxford University Press, Oxford, pp. 58-90, 2004.
[25] A.M. Turing, 'Intelligent Machinery', Internal report of National Physics Laboratory, 1948 (1948); reprinted in: J. Copeland (ed), *The essential Turing*, Oxford University Press, Oxford, pp. 410–432, 2004.
[26] A.M. Turing, 'Computing Machinery and Intelligence', *Mind*, **59**, 433–460, (1950).

Could There be a Turing Test for Qualia?

Paul Schweizer[1]

Abstract. The paper examines the possibility of a Turing test designed to answer the question of whether a computational artefact is a genuine subject of conscious experience. Even given the severe epistemological difficulties surrounding the 'other minds problem' in philosophy, we nonetheless generally believe that other human beings are conscious. Hence Turing attempts to defend his original test (2T) in terms of operational parity with the evidence at our disposal in the case of attributing understanding and consciousness to other humans. Following this same line of reasoning, I argue that the conversation-based 2T is far too weak, and we must scale up to the full linguistic and robotic standards of the Total Turing Test (3T).

Within this framework, I deploy Block's distinction between Phenomenal-consciousness and Access-consciousness to argue that passing the 3T could at most provide a sufficient condition for concluding that the robot enjoys the latter but not the former. However, I then propose a variation on the 3T, adopting Dennett's method of 'heterophenomenology', to rigorously probe the robot's purported 'inner' qualitative experiences. If the robot could pass such a prolonged and intensive Qualia 3T (Q3T), then the purely behavioural evidence *would* seem to attain genuine parity with the human case. Although success at the Q3T would not supply definitive proof that the robot was genuinely a subject of Phenomenal-consciousness, given that the external evidence is now equivalent with the human case, apparently the only grounds for denying qualia would be appeal to difference of *internal* structure, either physical-physiological or functional-computational. In turn, both of these avenues are briefly examined.

1 INTRODUCTION

According to the widely embraced 'computational paradigm', which underpins cognitive science, Strong AI and various allied positions in the philosophy of mind, computation (of one sort or another) is held to provide the scientific key to explaining mentality in general and, ultimately, to reproducing it artificially. The paradigm maintains that cognitive processes are essentially computational processes, and hence that intelligence in the natural world arises when a material system implements the appropriate kind of computational formalism. So this broadly Computational Theory of Mind (CTM) holds that the mental states, properties and contents sustained by human beings are fundamentally computational in nature, and that computation, at least in principle, opens the possibility of creating artificial minds with comparable states, properties and contents.

Traditionally there are two basic features that are held to be essential to minds and which decisively distinguish mental from non-mental systems. One is representational content: mental states can be *about* external objects and states of affairs. The other is conscious experience: roughly and as a first approximation, there is *something it is like* to be a mind, to be a particular mental subject. As a case in point, there is something it is like for me to be consciously aware of typing this text into my desk top computer. Additionally, various states of my mind are concurrently directed towards a number of different external objects and states of affairs, such as the letters that appear on my monitor. In stark contrast, the table supporting my desk top computer is not a mental system: there are no states of the table that are properly about anything, and there is nothing it is like to be the table.

Just as it seems doubtful that the term 'mind' should be applied to a system with no representational states, so too, many would claim that a system entirely devoid of conscious experience cannot be a mind. Hence if the project of Strong AI is to be successful at its ultimate goal of producing a system that truly counts as an artificially engendered locus of mentality, then it would seem necessary that this computational artefact be fully conscious in a manner comparable to human beings.

2 CONSCIOUSNESS AND THE ORIGINAL TURING TEST

In 1950 Turing [1] famously proposed an answer to the question 'Can (or could) a machine think?' by replacing it with the more precise and empirically tractable question 'Can (or could) a machine pass a certain type of test?', which mode of assessment has since become universally referred to as the 'Turing test' (2T). In brief, (the standardized version of) Turing's test is an 'imitation game' involving three players: a computational artifact and two humans. One of the humans is the 'judge' and can pose questions to the remaining two players, where the goal of the game is for the questioner to determine which of the two respondents is the computer. If, after a set amount of time, the questioner guesses correctly, then the machine loses the game, and if the questioner is wrong then the machine wins. Turing claimed, as a basic theoretical point, that any machine that could win the game a suitable number of times has passed the test and should be judged to be intelligent, in the sense that its behavioral performance has been demonstrated to be indistinguishable from that of a human being.

In his prescient and ground breaking article, Turing explicitly considers the application of his test to the question of *machine consciousness*. This is in section (4) of the paper, where he considers the anticipated 'Argument from Consciousness' objection to the validity of his proposed standard for answering the question 'Can a machine think?'. The objection is that, as per the above, consciousness is a necessary precondition for genuine thinking and mentality, and that a machine might fool its

[1] Institute for Language, Cognition and Computation, School of Informatics, Univ. of Edinburgh, EH8 9AD, UK. Email: paul@inf.ed.ac.uk.

interlocutor and pass the purely behavioural 2T, and yet remain completely devoid of internal conscious experience. Hence merely passing the 2T does not provide a sufficient condition for concluding that the system in question possesses the characteristics required for intelligence and *bona fide* thinking. Hence the 2T is inherently defective.

Turing's defensive strategy is to invoke the well known and severe epistemological difficulties surrounding the very same question regarding our fellow human beings. This is the notorious 'other minds problem' in philosophy – how do you know that other people actually have a conscious inner life like your own? Perhaps everyone else is a zombie and you're the only conscious being in the universe. As Turing humorously notes, this type of 'solipsistic' view (although more accurately characterized as a form of other minds skepticism, rather than full blown solipsism), while logically impeccable, tends to make communication difficult, and rather than continually arguing over the point, it is usual to simply adopt the polite convention that everyone is conscious.

Turing notes that on its most extreme construal, the only way that one could be sure that a machine or another human being is conscious and hence genuinely thinking is to *be* the machine or the human and *feel oneself* thinking. In other words, one would have to gain first person access to *what it's like* to be the agent in question. And since this is not an empirical option, we can't know with certainty whether any other system is conscious – all we have to go on is behaviour. Hence Turing attempts to justify his behavioural test that a machine can think, and *ipso facto*, has conscious experience, by claiming parity with the evidence at our disposal in the case of other humans. He therefore presents his anticipated objector with the following dichotomy: either be guilty of an inconsistency by accepting the behavioural standard in the case of humans but not computers, or maintain consistency by rejecting it in *both* cases and embracing solipsism. He concludes that most consistent proponents of the argument from consciousness would chose to abandon their objection and accept his test rather than be forced into the solipsistic position.

However, it is worth applying some critical scrutiny to Turing's reasoning at this early juncture. Basically, he seems to be running *epistemological* issues together with *semantical* and/or *factive* questions which should properly be kept separate. It's one thing to ask what we *mean* by saying that a system has a mind – i.e. what essential traits and properties are we ascribing to it with the use of the term; while it's quite another thing to ask how we can *know* that a given system actually satisfies this meaning and hence really does have a mind. Turing's behaviouristic methodology has a strong tendency to collapse these two themes, but it is important to note that they are conceptually distinct. In the argument from consciousness, the point is that we *mean* something substantive, something more than just verbal stimulus-response patterns, when we attribute mentality to a system. In this case the claim is that we mean that the system in question has conscious experience, and this property is required for any agent to be accurately described with the term 'mind'.

So one could potentially hold that consciousness is essential to mentality (because that's part of the core meaning of the term) and that:

(1) other human beings are in fact conscious
(2) the computer is in fact unconscious
(3) therefore, the computer doesn't have a mind, even though it passes the 2T.

This could be the objective state of affairs that genuinely obtains in the world, and this is completely independent of whether we can *know*, with certainty, that premises (1) and (2) are actually true. Although epistemological and factive issues are intimately related and together inform our general practices and goals of inquiry, nonetheless we could still be correct in our assertion, without being able to *prove* it's correctness. So if one thought that consciousness was essential to genuine mentality, then one could seemingly deny that any purely behaviouristic standard was sufficient to test for whether a system had or was a mind.

In the case of other human beings, we certainly take behaviour as *evidence* that they are conscious, but the evidence could in principle overwhelmingly support a *false* conclusion, in both directions. For example, someone could be in a comatose state where they could show no evidence of being conscious because they could make no bodily responses. But in itself this wouldn't make them unconscious. They could still be cognizant of what was going on and perhaps be able to report, retrospectively, on past events once out of their coma. And again, maybe *some* people really are zombies, or sleepwalkers, and exhibit all the appropriate external signs of consciousness even though they're really asleep or under some voodoo spell - it's certainly a conceivable state of affairs which cannot simply be ruled out *a priori*.

Historically, there has been disagreement regarding the proper interpretation of Turing's position regarding the intended import of his test. Some have claimed that the 2T is proposed as an operational *definition* of intelligence, thinking, etc., (e.g. Block [2], French [3]), and as such it has immediate and fundamental faults. However, in the current discussion I will adopt a weaker reading and interpret the test as purporting to furnish an empirically specifiable criterion for when intelligence can be legitimately *ascribed* to an artefact. On this reading, the main role of behavior is inductive or evidential rather than constitutive, and so behavioral tests for mentality do not provide a necessary condition nor a reductive definition. At most, all that is warranted is a *positive* ascription of intelligence or mentality, *if* the test is adequate *and* the system passes. In the case of Turing's 1950 proposal, the adequacy of the test is defended almost entirely in terms of parity of input/output performance with human beings, and hence alleges to employ the same operational standards that we tacitly adopt when ascribing conscious thought processes to our fellow creatures.

Thus the issue would appear to hinge upon the degree of evidence a successful 2T performance provides for a positive conclusion in the case of a computational artefact, (i.e. for the negation of (2) above), and how this compares to the total body of evidence that we have in support of our belief in the truth of (1). We will only be guilty of an inconsistency or employing a double standard if the two are on a par and we nonetheless dogmatically still insist on the truth of both (1) and (2). But if it turns out to be the case that our evidence for (1) is significantly better than for the negation of (2), then we are not forced into Turing's dichotomy. And in terms of the original 2T, I think there is clearly very little parity with the human case. We rely on far more than simply *verbal* behaviour in arriving at the polite convention that other human beings are conscious. In addition to conversational data, we lean very heavily on their bodily actions involving perception of the spatial environment, navigation,

physical interaction, verbal and other modes of response to communally accessible non-verbal stimuli in the shared physical surroundings, etc. So the purely conversational standards of the 2T are not nearly enough to support a claim of operational parity with humans. In light of the foregoing observations, in order to move towards evidential equivalence in terms of observable behaviour, it is necessary to break out of the closed syntactic bubble of the 2T and scale up to a full linguistic *and robotic* version of the test. But before exploring this vastly strengthened variation as a potential test for the presence of conscious experience in computational artefacts, in the next section I will briefly examine the notion of consciousness itself, since we first need to attain some clarification regarding the phenomenon in question, before we go looking for it in robots.

3 TWO TYPES OF CONSCIOUSNESS

Even in the familiar human case, consciousness is a notoriously elusive phenomenon, and is quite difficult to characterize rigorously. In addition, the word 'consciousness' is not used in a uniform and univocal manner, but rather appears to have different meanings in different contexts of use and across diverse academic communities. Block [4] provides a potentially illuminating philosophical analysis of the distinction and possible relationship between two common uses of the word. Block contends that consciousness is a 'mongrel' term connoting a number of different concepts and denoting a number of different phenomena. He attempts to clarify the issue by distinguishing two basic and distinct forms of consciousness that are often conflated: *Phenomenal* or P-consciousness and *Access* or A-consciousness. Very roughly, "Phenomenal consciousness is experience: what makes a state phenomenally conscious is that there is 'something it's like' to be in that state". Somewhat more controversially, Block holds that P-conscious properties, as such, are "distinct from any cognitive, intentional or functional property." The notoriously difficult explanatory gap problem in philosophical theorizing concerns P-consciousness – e.g. how is it possible that appeal to a physical brain process could explain what it is like to see something as red?

So we must take care to distinguish this type of purely qualitative, Phenomenal consciousness, from Access consciousness, the latter of which Block sees as an *information processing* correlate of P-consciousness. A-consciousness states and structures are those which are directly available for control of speech, reasoning and action. Hence Block's rendition of A-consciousness is similar to Baars' [5] notion that conscious *representations* are those that are broadcast in a global workspace. The functional/computational approach holds that the level of analysis relevant for understanding the mind is one that allows for multiple realization, so that in principle the same mental states and phenomena can occur in vastly different types of physical systems which implement the same abstract functional or computational structure. As a consequence, a staunch adherent of the functional-computational approach is committed to the view that the same *conscious* states must be preserved across widely diverse type of physical implementation. In contrast, a more 'biological' approach holds that details of the particular physical/physiological realization matter in the case of conscious states. Block says that if P = A, then the information processing side is right, while if the biological nature of experience is crucial then we can expect that P and A will diverge.

A crude difference between the two in terms of overall characterization is that P-consciousness content is qualitative while A-consciousness content is representational. A-conscious states are necessarily transitive or intentionally directed, they are always states of consciousness *of*. However. P-conscious states don't have to be transitive. On Block's account, the paradigm P-conscious states are the qualia associated with sensations, while the paradigm A-conscious states are propositional attitudes. He maintains that the A-type is nonetheless a genuine form of consciousness, and tends to be what people in cognitive neuroscience have in mind, while philosophers are traditionally more concerned with qualia and P-consciousness, as in the hard problem and the explanatory gap. In turn, this difference in meaning can lead to mutual misunderstanding. In the following discussion I will examine the consequences of the distinction between these two types of consciousness on the prospects of a Turing test for consciousness in artefacts.

4 THE TOTAL TURING TEST

In order to attain operational parity with the evidence at our command in the case of human beings, a Turing test for even basic linguistic understanding and intelligence, let alone conscious experience, must go far beyond Turing's original proposal. The conversational 2T relies solely on verbal input/output patterns, and these alone are not sufficient to evince a correct *interpretation* of the manipulated strings. Language is primarily about *extra-linguistic* entities and states of affairs, and there is nothing in a cunningly designed program for pure syntax manipulation which allows it to break free of this closed loop of symbols and demonstrate a proper correlation between word and object. When it comes to judging human language users in normal contexts, we rely on a far richer domain of evidence. Even when the primary focus of investigation is language proficiency and comprehension, sheer *linguistic* input/output data is not enough. Turing's original test is not a sufficient condition for concluding that the computer genuinely understands or refers to anything with the strings of symbols it produces, because the computer doesn't have the right sort of relations and interactions with the objects and states of affairs *in the real world* that its words are supposed to be about. To illustrate the point; if the computer has no eyes, no hands, no mouth, and has never seen or eaten anything, then it is not talking about hamburgers when its program generates the string of English symbols 'h-a-m-b-u-r-g-e-r-s' – it's merely operating inside a closed loop of syntax.

In sharp contrast, *our* talk of hamburgers is intimately connected to *nonverbal* transactions with the objects of reference. There are 'language entry rules' taking us from nonverbal stimuli to appropriate linguistic behaviours. When given the visual stimulus of being presented with a pizza, a taco and a kebab, we can produce the salient utterance "Those particular foodstuffs are not hamburgers". And there are 'language exit rules' taking us from linguistic expressions to appropriate nonverbal actions. For example, we can follow complex verbal instructions and produce the indicated patterns of behaviour, such as finding the nearest Burger King on the basis of a description of its location in spoken English. Mastery of both of these types of rules is essential for deeming that a

human agent understands natural language and is using expressions in a correct and referential manner - and the hapless 2T computer *lacks* both.[2]

And when it comes to testing for conscious experience, we again need these basic additional dimensions of perception and action *in the real world* as an essential precondition. The fundamental limitations of mere conversational performance naturally suggest a strengthening of the 2T, later named the Total Turing Test (3T) by Harnad [7], wherein the repertoire of relevant behaviour is expanded to include the full range of intelligent human activities. This will require that the computational procedures respond to and control not simply a teletype system for written inputs and outputs, but rather a well crafted artificial body. Thus in the 3T the scrutinized artefact is a *robot*, and the data to be tested coincide with the full spectrum of behaviours of which human beings are normally capable. In order to succeed, the 3T candidate must be able to do, in the real world of objects and people, everything that intelligent people can do. Thus Harnad expresses a widely held view when he claims that the 3T is "...no less (nor more) exacting a test of having a mind than the means we already use with one another... [and, echoing Turing] there is no stronger test, short of *being* the candidate". And, as noted above, the latter state of affairs is not an empirical option. examined.[3]

Since the 3T requires the ability to perceive and act in the real world, and since A-consciousness states and structures are those which are directly available for control of speech, reasoning and action, it would seem to follow that the successful 3T robot must be A-conscious. For example, in order to pass the test, the robot would have to behave in an appropriate manner in any number of different scenarios such as the following. The robot is handed a silver platter on which a banana, a boiled egg, a teapot and a hamburger are laid out. The robot is asked to pick up the piece of fruit and throw it out the window. Clearly the robot could not perform the indicated action unless it had direct information processing access to the identity *of* the salient object, its spatial location, the movements *of* its own mechanical arm, the location and geometrical properties *of* the window, etc. Such transitive, intentionally directed A-conscious states are plainly required for the robot to pass the test.

But does it follow that the successful 3T robot is P-conscious? It seems, not, since on the face of it there appears to be no reason why the robot could not pass the test relying on A-consciousness alone. All that is being tested is its executive control of the cognitive processes enabling it to reason correctly and perform appropriate verbal and bodily actions in response to a myriad of linguistic and perceptual inputs. These abilities are demonstrated solely through its external behaviour, and so far, there seems to be no reason for P-conscious states to be invoked. Since the 3T is primarily intended to test the robot's overall intelligence and linguistic understanding in the actual world, the A-conscious robot could conceivably pass the 3T while at the same time there *is nothing it is like* to be the 3T robot passing the test. We are now bordering on issues involved in demarcating the 'easy' from the 'hard' problems of consciousness, which, if pursued at this point, would be moving in a direction not immediately relevant to the topic at hand. So rather than exploring arguments relating to this deeper theme, I will simply contend that passing the 3T provides a sufficient condition for Block's version of A-consciousness, but not for P-consciousness, since it could presumably be passed by an artefact devoid of qualia.

Many critics of Block's basic type of view (including Searle [9] and Burge [10]) argue that if there can be such functional 'zombies' that are A-conscious but not P-conscious, then they are not genuinely conscious *at all*. Instead, A-consciousness is better characterized as a type of 'awareness', and is a form of consciousness only to the extent that it is parasitic upon P-conscious states. So we could potentially have a 3T for A-consciousness, but then the pivotal question arises, is A-consciousness without associated qualitative presentations really a form of *consciousness*? Again, I will not delve into this deeper and controversial issue in the present discussion, but simply maintain that the successful 3T robot does at least exhibit the type of A-*awareness* that people in, e.g., cognitive neuroscience tend to call consciousness. But as stated earlier, 'consciousness' is a multifaceted term, and there are also good reasons for *not* calling mere A-awareness without qualia a full-fledged form of consciousness.

For example, someone who was drugged or talking in their sleep could conceivably pass the 2T while still 'unconscious', that is A-'conscious' but not P-conscious. And a human sleep walker might even be able to pass the verbal and robotic 3T while 'unconscious' (again A-'conscious' but not P-conscious). What this seems to indicate is that only A-'consciousness' can be positively ascertained by behaviour. But there is an element of definitiveness here, since it seems plausible to say that an agent *could not* pass the 3T without being A-'conscious', at least in the minimal sense of A-awareness. If the robot were warned 'mind the banana peel' and it was *not* A-aware of the treacherous object in question on the ground before it, emitting the frequencies of electromagnetic radiation appropriate for 'banana-yellow', then it would not deliberately step over the object, but rather would slip and fall and fail the test.

5 A TOTAL TURING TEST FOR QUALIA

In the remainder of the paper I will not pursue the controversial issue as to whether associated P-consciousness is a necessary condition for concluding that the A-awareness of the successful 3T robot is genuinely a form of consciousness *at all*. Instead, I will explore an intensification of the standard 3T intended to prod more rigorously for *evidential support* of the presence of P-conscious states. This Total Turing Test for qualia (Q3T) is a more focused scrutiny of the successful 3T robot which emphasizes rigorous and extended verbal and descriptive probing into the qualitative aspects of the robot's purported internal experiences. So the Q3T involves unremitting questioning and verbal analysis of the robot's qualitative inner experiences, in reaction to a virtually limitless variety of salient external stimuli, such as paintings, sunsets, musical

[2] Shieber [6] provides a valiant and intriguing rehabilitation/defense of the 2T, but it nonetheless still neglects crucial data, such as mastery of language exit and entry rules. Ultimately Shieber's rehabilitation in terms of interactive proof requires acceptance of the notion that *conversational* input/response patters alone are sufficient, which premise I would deny for the reasons given. The program is still operating within a closed syntactic bubble.

[3] See Schweizer [8] for an argument to the effect that even the combined linguistic and robotic 3T is still too weak as a definitive *behavioural* test of artificial intelligence.

performances, tastes, textures, smells, pleasures and pains, emotive reactions...

Turing suggests a precursor version of this strategy in his 1950 discussion of the argument from consciousness, where he observes that the question of machine consciousness could be addressed by a sustained *viva voce*, where the artefact was asked questions directly concerning its aesthetic and other types of qualitative reactions and judgement in response to opened-ended questioning by the interrogator. Turing provides a conjectural illustration of the method in the guise of a 'sonnet writing' programme being quizzed by a human judge.

> Interrogator: In the first line of your sonnet which reads "Shall I compare thee to a summer's day," would not "a spring day" do as well or better?
>
> Witness: It wouldn't scan.
>
> Interrogator: How about "a winter's day," that would scan all right.
>
> Witness: Yes, but nobody wants to be compared to a winter's day.
>
> Interrogator: Would you say Mr. Pickwick reminded you of Christmas?
>
> Witness: In a way.
>
> Interrogator: Yet Christmas is a winter's day, and I do not think Mr. Pickwick would mind the comparison.
>
> Witness: I don't think you're serious. By a winter's day one means a typical winter's day, rather than a special one like Christmas.
>
> And so on....

The above sample dialogue serves as a good reminder of just how difficult the original 2T really is (and consequently why it hasn't yet been passed). However, this conjectured scenario is still confined to a merely conversational setting of verbal inputs and verbal outputs, and hence falls far short of the behavioural potential of a full 3T edition, as well as the concomitant evidential standards applicable in the human case.

Plebe and Perconti [11] put forward a strengthened adaptation of a 2T-style *viva voce*, where the pivotal difference is that, in addition to merely linguistic inputs, the computer must now give appropriate and testable conversational reactions to *uploaded images*. This is an interesting and important augmentation of the original 2T, since the inputs are no longer strictly linguistic, and the test is aimed at evaluating verbally plausible responses to stimuli that, *to us at least*, have a phenomenal aspect. As an example of the method, Plebe and Perconti supply an excerpt from a hypothetical conversation.

> Interrogator: Do you want to look at a picture of me?
>
> Machine: Oh yes, thanks, let's upload that.
> <.... uploading>
>
> Machine: Mmmh, I see several people here, who are you?
>
> Interrogator: Try to guess.
>
> Machine: Well, I know you are blond and shy, so I would guess the second from the left.

This appears to be an order of magnitude jump over the purely verbal 2T, and accordingly its standards of satisfaction are even more futuristic. However, in terms of the ultimate goal of providing a test, the passing of which constitutes a sufficient condition for the presence of genuine conscious experience in an artefact, it should be noted that the inputs, at a crucial level of analysis, remain purely syntactic and non-qualitative, in that the uploaded image must take the form of a digital file. Hence this could at most provide evidence of some sort of (proto) A-awareness in terms of salient data extraction and attendant linguistic conversion from a digital source, where the phenomenal aspects produced in humans by the original (pre-digitalized) image are systematically corroborated by the computer's linguistic outputs when responding to the inputted code.

Although a major step forward in terms of expanding the input repertoire under investigation, as well as possessing the virtue of being closer to the limits of practicality in the nearer term future, this proposed new qualia 2T still falls short of the full linguistic and robotic Q3T. In particular it tests, in a relatively limited manner, only one sensory modality, and in principle there is no reason why this method of scrutiny should be restricted to the intake of photographic images represented in digital form. Hence a natural progression would be to test a computer on uploaded audio files as well. However, this expanded 2T format is still essentially passive in nature, where the neat and tidy uploaded files are hand fed into the computer by the human interrogator, and the outputs are confined to mere verbal response. Active perception of and reaction to distal objects in the real world arena are critically absent from this test, and so it fails to provide anything like evidential parity with the human case. And given the fact that the selected non-linguistic inputs take the form of digitalized representations of possible visual (and/or auditory) stimuli, there is still no reason to think that there is anything it is like to be the 2T computer processing the uploaded encoding of an image of, say, a vivid red rose.

But elevated to a full 3T arena of shared external stimuli and attendant discussion and analysis, the positive evidence of a victorious computational artefact would become exceptionally strong indeed. So the extended Q3T is based on a methodology akin to Dennett's [12] 'heterophenomenology' - given the robot's presumed success at the standard Total Turing Test, we count this as behavioural evidence sufficient to warrant the application of the intentional stance, wherein the robot is treated as a rational agent harbouring beliefs, desires and various other mental states exhibiting intentionality, and who's actions can be explained and predicted on the basis of the content of these states. Accordingly, the robot's salient sonic emissions are *interpreted* as natural language utterances asserting various propositions and expressing assorted contents. For the reasons delineated above in section 4, I would argue that this interpretive step and application of the intentional stance to a mere artefact is not evidentially warranted in a limited 2T type of setting, and

hence in the latter case, the syntactic tokens emitted by the computer cannot be given the semantic value that we would attribute to utterances produced by our fellow human beings.

However, following Dennett, the gathering of scientifically rigorous empirical evidence proceeds with neutrality regarding the possible conscious experience of the robot. In this extended Q3T context, the artefact is queried not about uploaded images as 'qualitative' *input* data, but rather about *its own* subjective reactions and attendant introspective states. The robot is not hand fed selected digital files, but rather is presented with the same external stimuli as its human interlocutor, the latter of whom can then interrogate the robot regarding its purported qualitative experiences in response to the shared stimuli. Hence the human and the robot can both observe the same Dürer and the same Dali paintings, listen to the same Vivaldi and Schubert concertos, attend the same wine tasting event, eat the same brand of hamburger... And the robot must discourse on its subjective impressions and reactions like a true aesthete, and where this discourse can be calibrated by the human with respect to the latter's own qualitative reactions.

As an example of the method, an excerpt from a hypothetical Q3T interaction might look something like the following.

> Interrogator: Isn't the sunset over the Pacific magnificent?
>
> Robot: Yes, the colours are sublime. Let's go to the top of that rise for a better view...
> Excellent, now we have the full panorama.
>
> Interrogator: What impression does it make on you?
>
> Robot: As the glowing orange orb of the sun begins to touch the calm blue line of the horizon, yet another day is passing. Now the radiant disc begins to sink into the sea and the background crimson intensifies and grows warmer. It gives me a feeling of sadness, but at the same time a glimpse of ineffable beauty...
> Ouch!! Why the hell did you pinch me?
>
> Interrogator: Just wanted to bring you back down to earth - you were starting to wax a bit too lyrical.
>
> Robot: Thanks very much for your concern. Could you please pass me the bottle of delicious chilled Chardonnay - I want to savour another glass along with the last rays of the setting sun.
>
> Interrogator: Here you go.
>
> Robot: Arrrgh, that tastes disgusting! - what happened to the wine?
>
> Interrogator: Uhh, I just stirred in a little marmite when you weren't looking - wanted to see how you'd react. This is a Q3T, after all...

Even though a merely A-conscious robot could conceivably pass the verbal and robotic 3T while at the same time as there being *nothing it is like* for the robot passing the test, in this more focussed version of the 3T the robot would at least have to be able to go on at endless length *talking about* what it's like. And this talk must be in response to an open ended range of different combinations of sensory inputs, which are shared and monitored by the human judge. Such a test would be both subtle and extremely demanding, and it would be nothing short of remarkable if it could *not* detect a fake. And presumably a human sleepwalker who could pass a normal 3T as above would nonetheless *fail* this type of penetrating Q3T (or else wake up in the middle!), and it would be precisely on the grounds of such failure that we would infer that the human was actually asleep and not genuinely P-conscious of what was going on.

If sufficiently rigorous and extended, this would provide extremely powerful inductive evidence, and indeed to pass the Q3T the robot would have to attain full evidential parity with the human case, in terms of externally manifested behaviour.

6 BEYOND BEHAVIOUR

So on what grounds might one *consistently deny* qualitative states and P-consciousness in the case of the successful Q3T robot and yet grant it in the case of a behaviourally indistinguishable human? The two most plausible considerations that suggest themselves are both based on an appeal to essential differences of *internal* structure, either physical/physiological or functional/computational. Concerning the latter case, many versions of CTM focus solely on the functional analysis of propositional attitude states such as belief and desire, and simply ignore other aspects of the mind, most notably consciousness and qualitative experience. However others, such as Lycan [13], try to extend the reach of Strong AI and the computational paradigm, and contend that *conscious states* arise via the implementation of the appropriate computational formalism. Let us denote this extension of the basic CTM framework to the explanation of conscious experience 'CTM+'. And a specialized version of CTM+ might hold that qualitative experiences arise in virtue of the particular functional and information processing structure of the *human* brand of cognitive architecture, and hence that, even though the robot is indistinguishable in terms of input/output profiles, nonetheless its internal processing structure is sufficiently different from ours to block the inference to P-consciousness. So the non-identity of abstract functional or computational structure might be taken to undermine the claim that bare behavioural equivalence provides a sufficient condition for the presence of internal conscious phenomena.

At this juncture, the proponent of artificial consciousness might appeal to a version of Van Gulick's [14] defense of functionalism against assorted 'missing qualia' objections. When aimed against functionalism, the missing qualia arguments generally assume a deviant realization of the very same abstract computational procedures underlying human mental phenomena, in a world that's nomologically the same as ours in all respects, and the position being supported is that consciousness is to be equated with states of the biological brain, rather than with any *arbitrary* physical state playing the same functional role as a conscious brain process. For example, in Block's [15] well known 'Chinese Nation' scenario, we are asked to imagine a case where each person in China plays the role of a neuron in the human brain and for some (rather brief) span of time the entire nation cooperates to implement the same

computational procedures as a conscious human brain. The rather compelling 'common sense' conclusion is that even though the entire Chinese population may implement the same computational structure as a conscious brain, there are nonetheless no purely qualitative conscious states in this scenario outside the conscious Chinese *individuals* involved. And this is then taken as a counterexample to purely functionalist theories of consciousness.

Van Gulick's particular counter-strategy is to claim that the missing qualia argument begs the question at issue. How do we know, *a priori*, that the very same functional role *could* be played by arbitrary physical states that were unconscious? The anti-functionalist seems to beg the question by assuming that such deviant realizations are possible in the first place. At this point, the burden of proof may then rest on the functionalist to try and establish that there are in fact functional roles in the human cognitive system that could only be filled by *conscious* processing states. Indeed, this strategy seems more interesting than the more dogmatic functionalist line that isomorphism of abstract functional role *alone* guarantees the consciousness of any physical state that happens to implement it.

So to pursue this strategy, Van Gulick examines the psychological roles played by phenomenal states in humans and identifies various cognitive abilities which *seem* to require both conscious and self-conscious awareness – e.g. abilities which involve reflexive and meta-cognitive levels of representation. These include things like planning a future course of action, control of plan execution, acquiring new non-habitual task behaviours These and related features of human psychological organization seem to require a conscious self-model. In this manner, conscious experience appears to play a *unique* functional role in broadcasting 'semantically transparent' information throughout the brain. In turn, the proponent of artificial consciousness might plausibly claim that the successful Q3T robot must possesses analogous processing structures in order to evince the equivalent behavioural profiles when passing the test. So even though the processing structure might not be identical to that of human cognitive architecture, it must nonetheless have the same basic cognitive abilities as humans in order to pass the Q3T, and if these processing roles in humans require phenomenal states, then the robot must enjoy them as well.

However, it is relevant to note that Van Gulick's analysis seems to blur Block's distinction between P-consciousness and A-consciousness, and an obvious rejoinder at this point would be that all of the above processing roles in both humans and robots could in principle take place with only the latter and not the former. Even meta-cognitive and 'conscious' self models could be accounted for merely in terms of A-*awareness*. And this brings us back to the same claim as in the standard 3T scenario - that even the success of the Q3T robot could conceivably be explained without invoking P-consciousness *per se*, and so it still fails as a sufficient condition for attributing full blown qualia to computational artefacts.

7 MATTER AND CONSCIOUSNESS

Hence functional/computational considerations seem too weak to ground a positive conclusion, and this naturally leads to the question of the physical/physiological status of qualia. If even meta-cognitive and 'conscious' self models in humans could in principle be accounted for merely in terms of A-*awareness*, then how and why do *humans* have purely qualitative experience? One possible answer could be that P-conscious states are essentially *physically based phenomena*, and hence result from or supervene upon the particular structure and causal powers of the actual central nervous system. And this perspective is re-enforced by what I would argue (on the following independent grounds) is the fundamental inability of abstract functional role to provide an adequate theoretical foundation for qualitative experience.

Unlike computational formalisms, conscious states are inherently *non-abstract*; they are *actual*, occurrent phenomena extended in physical time. Given multiple realizability as a hallmark of the theory, CTM+ is committed to the result that qualitatively identical conscious states are maintained across widely different kinds of physical realization. And this is tantamount to the claim that an actual, substantive and *invariant* qualitative phenomenon is preserved over radically diverse real systems, while at the same time, *no* internal physical regularities need to be preserved. But then there is no actual, occurrent factor which could serve as the causal substrate or supervenience base for the substantive and invariant phenomenon of internal conscious experience. The advocate of CTM+ cannot rejoin that it is *formal role* which supplies this basis, since formal role is abstract, and such abstract features can only be *instantiated* via actual properties, but they do not have the power to *produce* them.

The only (possible) non-abstract effects that instantiated formalisms are required to preserve must be specified in terms of their input/output profiles, and thus *internal* experiences, qua actual events, are in principle omitted. So (as I've also been argued elsewhere: see Schweizer [16,17]) it would appear that the non-abstract, occurrent nature of conscious states entails that they must depend upon intrinsic properties of the brain as a proper subsystem of the actual world (on the crucial assumption of *physicalism* as one's basic metaphysical stance - obviously other choices, such as some variety of dualism, are theoretical alternatives). It is worth noting that from this it *does not follow* that other types of physical subsystem could not share the relevant intrinsic properties and hence also support conscious states. It only follows that they would have this power in virtue of their intrinsic physical properties and *not* in virtue of being interpretable as implementing the same abstract computational procedure.

8 CONCLUSION

We know by direct first person access that the human central nervous system is capable of sustaining the rich and varied field of qualitative presentations associated with our normal cognitive activities. And it certainly *seems as if* these presentations play a vital role in our mental lives. However, given the above critical observation regarding Van Gulick's position, *viz.*, that all of the salient processing roles in *both* humans and robots could in principle take place strictly in terms of A-awareness without P-consciousness, it seems that P-conscious states are not actually necessary for explaining observable human behaviour and the attendant cognitive processes. In this respect, qualia are rendered *functionally* epiphenomenal, since purely qualitative states per se are not strictly required for a functional/computational account of human mentality. However, this is not to say that they are

physically epiphenomenal as well, since it doesn't thereby follow that this aspect of physical/physiological structure does not in fact play a causal role in the particular *human* implementation of this functional cognitive architecture. Hence it becomes a purely contingent truth that humans have associated P-conscious experience.

And this should not be too surprising a conclusion, on the view that the human mind is the product of a long course of exceedingly happenstance biological evolution. On such a view, perhaps natural selection has simply *recruited* this available biological resource to play vital functional roles, which in principle could have instead been played by P-*unconscious* but A-aware states in a *different type* of realization. And in this case, P-conscious states in humans are thus a form of 'phenomenal overkill', and nature has simply been an opportunist in exploiting biological vehicles that happened to be on hand, to play a role that could have been played by a more streamlined and less rich type of state, but where a 'cheaper' alternative was simply not available at the critical point in time. Evolution and natural selection are severely curtailed in this respect, since the basic ingredients and materials available to work with are a result of random mutation on existing precursor structures present in the organism(s) in question. And perhaps human computer scientists and engineers, not limited by what happens to get thrown up by random genetic mutations, have designed the successful Q3T robot utilizing a cheaper, artificial alternative to the overly rich biological structures sustained in humans.

So in the case of the robot, it would remain an open question whether or not the physical substrate underlying the artefact's cognitive processes had the requisite causal powers or intrinsic natural characteristics to sustain P-conscious states. Mere behavioural evidence on its own would not be sufficient to adjudicate, and an independent standard or criterion would be required.[4] So if P-conscious states are thought to be essentially physically based, for the reasons given above, and if the robot's Q3T success could in principle be explained through appeal to mere A-aware stets on their own, then it follows that the non-identity of the artefact's physical structure would allow one to consistently extend Turing's polite convention to one's conspecifics and yet withhold it from the Q3T robot.

REFERENCES

[1] A. Turing, 'Computing machinery and intelligence', *Mind* 59: 433-460 (1950).

[2] N. Block, 'Psychologism and behaviorism', *Philosophical Review* 90: 5-43 (1981).

[3] R. French, 'The Turing test: the first 50 years', *Trends in Cognitive Sciences* 4: 115-122 (2000).

[4] N. Block, 'On a confusion about a function of consciousness', *Behavioral and Brain Sciences* 18, 227-247, (1995).

[5] B. Baars, *A Cognitive Theory of Consciousness*, Cambridge University Press, (1988).

[6] S. Shieber, 'The Turing test as interactive proof', *Nous* 41:33-60 (2007).

[7] S. Harnad 'Other bodies, other minds: A machine incarnation of an old philosophical problem', *Minds and Machines* 1: 43-54, (1991).

[8] P. Schweizer, 'The externalist foundations of a truly total Turing test', *Mind & Machines*, DOI 10.1007/s11023-012-9272-4, (2012).

[9] J. Searle, *The Rediscovery of the Mind*, MIT Press, (1992).

[10] T. Burge, 'Two kinds of consciousness', in N. Block et al. (eds), *The Nature of Consciousness: Philosophical Debates*, MIT Press, (1997).

[11] A. Plebe and P. Perconti, 'Qualia Turing test: Designing a test for the phenomenal mind', in Proceedings of the First International Symposium *Towards a Comprehensive Intelligence Test (TCIT), Reconsidering the Turing Test for the 21st Century*, 16-19, (2010).

[12] D. Dennett, *Consciousness Explained*, Back Bay Books, (1992).

[13] W. G., Lycan, *Consciousness*, MIT Press, (1987).

[14] R. Van Gulick, 'Understanding the phenomenal mind: Are we all just armadillos?', in *Consciousness: Psychological and Philosophical Essays*, M. Davies and G. Humphreys (eds.), Blackwell, (1993).

[15] N. Block, 'Troubles with functionalism', in C. W. Savage (ed), *Perception and Cognition*, University of Minnesota Press, (1978).

[16] P. Schweizer, 'Consciousness and computation.' *Minds and Machines*, 12, 143-144, (2002)

[17] P. Schweizer, 'Physical instantiation and the propositional attitudes', *Cognitive Computation*, DOI 10.1007/s12559-012-9134-7, (2012).

[4] This highlights one of the intrinsic limitations of the Turing test approach to such questions, since the test is designed as an *imitation game*, and humans are the ersatz target. Hence the Q3T robot is designed to behave as if it had subjective, qualitative inner experiences indistinguishable from those of a human. However, if human qualia are the products of our particular internal structure (either physical-physiological or functional-computational), and if the robot is significantly different in this respect, then the possibility is open that the robot might be P-conscious and yet fail the test, simply because its resulting qualitative experiences are significantly different than ours. And indeed, a possibility in the reverse direction is that the robot might even *pass* the test and sustain an entirely different phenomenology, but where this internal difference is not manifested in its external behaviour.

Jazz and Machine Consciousness: Towards a New Turing Test

Antonio Chella[1] **and Riccardo Manzotti**[2]

Abstract. A form of Turing test is proposed and based on the capability for an agent to produce jazz improvisations at the same level of an expert jazz musician.
.

1 INTRODUCTION

The *Essay in the style of Douglas Hofstadter* [19] related to the system *EMI* by David Cope [11] [12], evokes a novel and different perspective for the Turing test. The main focus of the test should be creativity instead of linguistic capabilities: can a computer be so creative to the point that its creations could be indistinguishable from those of a human being?

According to Sterberg [36], creativity is the ability to produce something that is new and appropriate. The result of a creative process is not reducible to some sort of deterministic reasoning. No creative activity seems to identify a specific chain of activity, but an emerging *holistic* result [25].

Therefore, a creative agent should be able to generate novel artifacts not by following preprogrammed instructions, but on the contrary by means of a real creative act.

The problem of creativity has been widely debated in the field of automatic music composition. The previously cited *EMI* by David Cope, subject of the Hoftadter essay, produce impressive results: even for an experienced listener it is difficult to distinguish musical compositions created by these programs from those ones created by a human composer. There is no doubt that these systems capture some main aspects of the creative process, at least in music.

However, one may wonders if an agent can actually be creative without being conscious. In this regard, Damasio [14] suggests a close connection between consciousness and creativity. Cope himself in his recent book [13] discusses the relationship between consciousness and creativity. Although he does not take a clear position on this matter, he seem to favor the view according to which consciousness is not necessary for creative process. In facts, Cope asks if a creative agent should need to be aware of being creating something and if it needs to experience the results of its own creations.

The argument of consciousness is typically adopted [3] to support the thesis according to which an artificial agent can never be conscious and therefore it can never be really creative.

But recently, there has been a growing interest in machine consciousness [8] [9], i.e., the study of consciousness through the design and implementation of *conscious* artificial systems.

This interest is motivated by the belief that this new approach, based on the construction of *conscious* artifacts, can shed new light on the many critical aspects that affect the mainstream studies of consciousness from philosophy and neuroscience. Creativity is just one of these critical issues.

The relationship between consciousness and creativity is difficult and complex. On the one side some authors claim the need of awareness of the creative act. On the other side, it is suspected that many cognitive processes that are necessary for the creative act may happen in the absence of consciousness. However it is undeniable that consciousness is closely linked with the broader unpredictable and less *automatic* forms of cognition, like creativity.

In addition, we could distinguish between the mere production of new combinations and the aware creation of new content: if the wind would create (like the monkeys on a keyboard) a melody which is indistinguishable from the "Va Pensiero" by Giuseppe Verdi, it would be a creative act? Many authors would debate this argument [15].

In the following, we discuss some of the main features for a conscious agent like *embodiment*, *situatedness*, *emotions* and the capability to have conscious *experience*. These features will be discussed with reference to musical expression, and in particular to a specific form of creative musical expression, namely jazz improvisation. Musical expression seems to be a form of artistic expression that most of others is able to immediately produce conscious experience without filters. Moreover, differently from olfactory or tactile experiences, musical experience is a kind of structured experience.

According to Johnson-Laird [20], jazz improvisation is a specific form of expertise of great interest for the study of the mind. Furthermore, jazz is a particularly interesting case of study in relation to creativity. Creativity in a jazz musician is very different from typical models of creativity. In fact, the creativity process is often studied with regards to the production of new abstract ideas, as for example the creation of a new mathematical theory after weeks of great concentration. On the contrary, jazz improvisation is a form of immediate and continuous lively creation process which is closely connected with the external world made up of musical instruments, people, moving bodies, environments, audience and the other musicians.

2 CREATIVITY

There are at least two aspects of creativity that is worth distinguishing since the beginning: *syntactic* and *semantic* creativity. The first one is the capability to recombine a set of symbols according to various styles. In this sense, if we have enough patience and time, a random generator will create all the books of the literary world (but without understanding their meaning). The second aspect is the capability to generate new meaning that will be then *dressed* by appropriate symbols. These two aspects correspond to a good approximation to the etymological difference between the terms *intelligence* and

[1] University of Palermo, Italy, email: `antonio.chella@unipa.it`
[2] IULM University, Milan, Italy, email: `riccardo.manzotti@iulm.it`

intuition. Intelligence is often defined as the ability to find novel connections for different entities, but intuition should be able to do something more, i.e., to bring in something that was previously unavailable.

In short, the syntactic manipulation of symbols may occur without consciousness, but creativity does not seem to be possible without consciousness.

Machine consciousness is not only a technological challenge, but a novel field of research that has scientific and technological issues, such as the relationship between information and meaning, the ability for an autonomous agent to choose its own goals and objectives, the sense of self for a robot, the capability to integrate information into a coherent whole, the nature of experience. Among these issues there is the capability, for an artificial agent, to create and to experience its own creations.

A common objection to machine consciousness emphasizes the fact that biological entities may have unique characteristics that cannot be reproduced in artifacts. If this objection is true, machine consciousness may not be feasible. However, this contrast between biological and artificial entities has often been over exaggerated, especially in relation to the problems of consciousness. So far, nobody was able to satisfactorily prove that the biological entities may have characteristics that can not be reproduced in artificial entities with respect to consciousness. In fact, at the a meeting on machine consciousness in 2001 at Cold Spring Harbor Laboratories, the conclusion from Koch [23] was that no known natural law prevents the existence of subjective experience in artifacts. On the other hand, living beings are subject to the laws of physics, and yet are conscious, able to be creative and to prove experience.

The contrast between classic AI (focused on manipulation of syntactic symbols) and machine consciousness (open to consider the semantic and phenomenal aspects of the mind) holds in all his strength in the case of creativity.

Is artistic improvisation - jazz improvisation in particular - a conscious process? This is an open question. The musicologist Gunther Schuller [33] emphasizes the fact that jazz improvisation affects consciousness at all levels, from the minimal to the highest one. It is a very particular kind of creative process.

Jazz improvisation has peculiar features that set it apart from the traditional classic improvisation [29]: as part of Western classical music, improvisation is a kind of *real time* composition with the same rules and patterns of classic composition. On the contrary, jazz improvisation is based on a specific set of patterns and elements. The melody, the rhythm (the *swing*), the chord progressions are some of the issues that need to be analyzed and studied with stylistic and aesthetic criteria different from those of Western classical music [10].

3 EMBODIMENT

Embodiment does not simply mean that an agent must have a physical body, but also and above all, that different cognitive functions are carried out by means of aspects of the body. The aspect of corporeality seems to be fundamental to the musical performance and not only for jazz improvisation. In this regard, Sundberg & Verrillo [38] analyzed the complex feedback that the body of a player receives during a live performance. In facts, auditory feedback is not sufficient to explain the characteristics of a performance. The movement of the hands on the instrument, the touch and the strength needed for the instrument to play, the vibrations of the instrument propagated through the fingers of the player, the vibration of the air perceived by the player's body, are all examples of feedback guiding the musician during a performance. The player receives at least two types of bodily feedback: through the receptors of the skin and through the receptors of the tendons and muscles. Todd [39] assumed a third feedback channel through the vestibular apparatus.

Making music is essentially a body activity [26]. Embodiment is fundamental to jazz improvisation: can an agent without a body, such as a software like EMI that runs on a mainframe, be able to improvise? Apparently not, because it would miss the bodily feedback channels described above. And, in fact, the results obtained by EMI in the version *Improvisation* are modest and based on ad hoc solutions. The same problem arises for consciousness: can a software that run on a mainframe be conscious?

It does not seem that embodiment is a sufficient condition for consciousness, but it may be a necessary condition. Basically, a cognitive entity must be embodied in a physical entity. However, it is necessary to deeply reflect about the concept of embodiment.

Trivially, a cognitive agent can not exist without a body; even AI expert systems are embodied in a computer which is a physical entity. On the other hand it is not enough to have a body for an agent in order to be not trivially embodied: the Honda ASIMO robot[3], considered the state of the art of today robotic technology, is an impressive humanoid robot but its performances are essentially based on a standard controller in which the behaviors are almost completely and carefully defined in advance by its designers.

In addition, biology gives us many examples of animals, such as the cockroaches, whose morphology is complex and that allows them to survive without cognitive abilities.

The notion of embodiment is therefore much more deep and complex than we usually think. Not only the fact that an agent might have a body equipped with sophisticated sensors and actuators, but other conditions must be met. The concept of embodiment requires the ability to appreciate and process the different feedback from the body, just like an artist during a musical live performance.

4 SITUATEDNESS

In addition to having a body, an agent is part of an environment, i.e., it is *situated*. An artist, during a jam session, is typically situated in a group where she has a continuous exchange of information. The artist receives and provides continuous feedback with the other players of the group, and sometimes even with the audience, in the case of live performances.

The classical view, often theorized in textbooks of jazz improvisation [10], suggests that during a session, the player follows his own musical path largely made up by a suitable musical sequence of previously learned patterns. This is a partial view of an effective jazz improvisation. Undoubtedly, the musician has a repertoire of musical patterns, but she is also able to deviate from its path depending on the feedback she receives from other musicians or the audience, for example from

[3] http://asimo.honda.com

suggestions from the rhythm section or due to signals of appreciation from the listeners.

Cognitive scientists (see, e.g., [20]) typically model jazz improvisation processes by means of Chomsky formal grammars. This kind of model appears problematic because it does not explain the complexity of the interaction between the player, the rest of the group and the audience.

A more accurate model should take into account the main results from *behavior-based* robotics [5]. According to this approach, a musician may use a repertoire of behaviors that are activated according to the input she receives and according to an appropriate priority based on her musical sensibility. Interesting experiments in this direction have been recently described in the literature. *Roboser* [27] is an autonomous robot that can move autonomously in an environment and generate sound events in real time according to its internal state and to the sensory input it receives from the environment. *EyesWeb* [6] is a complex system that analyzes body movements and gestures with particular reference to emotional connotations in order to accordingly generate sound and music in real time and also to suitably control robots.

Continuator [28] is a system based on a methodology similar to EMI, but differently from it, is able to learn and communicate in real time with the musician. For example, the musician suggests that musical phrases and the system is able to learn the style of the musician and to continue and complete the sentences by interacting with the musician.

However, the concept of situated agent, as the concept of embodiment, is a complex and articulate one. An effective situated agent should develop a tight integration development with their surrounding environment so that, like a living being, its body structure and cognition would be the result of a continuous and constant interaction with the external environment.

A true situated agent is an agent that absorbs from its surroundings, changes according to it and, in turn, it changes the environment itself. A similar process occurs in the course of jazz improvisation: the musicians improvise on the basis of their musical and life experiences accumulated and absorbed over the years. The improvisation is then based on the interaction and also, in the case of a jazz group, even of past interactions with the rest of the group. Improvisation is modified on the basis of suggestions received from other musicians and audience, and in turn changes the performances of the other group musicians. A good jazz improvisation is an activity that requires a deeply situated agent.

5 EMOTIONS

Many scholars consider emotions as a basic element for consciousness. Damasio [14] believes that emotions form a sort of proto-consciousness upon which higher forms of consciousness are developed. In turn, consciousness, according to this frame of reference, is intimately related with creativity.

The relationships between emotions and music have been widely analyzed in the literature, suggesting a variety of computational models describing the main mechanisms underlying the evocation of emotions while listening to music [21] [22].

In the case of a live performance as a jazz improvisation, the link between music and emotions is a deep one: during a successful performance the player create a tight *empathic* relationship between herself and the listeners.

Gabrielsson & Juslin [17] conducted an empirical analysis of the emotional relationship between a musician and the listeners. According to this analysis, a song arouses emotions on the basis of its structure: for example, a *sad* song is in a minor key, it has a slow rhythm and the dissonances are frequent, while an *exciting* song is fast, strong, with few dissonances.

The emotional intentions of a musician during a live performance can be felt by the listener with greater or lesser effectiveness depending on the song itself. The basic emotional connotations such as the joy or the sadness are easier to transmit, while more complex connotation such as *solemnity* are more difficult to convey. The particular musical instrument employed has a relevance in the communication of emotions, and of course the degree of achieved empathy depends on the skill of the performer. This analysis shows that an agent, to make an effective performance, must be able to convey emotions and to have a model (even implicit) of them.

This hypothesis is certainly attractive, but it is unclear how to translate it into computational terms. So far, many computational models of emotions have been proposed in the literature. This is a very prolific field of research for robotics [16].

However, artificial emotions have been primarily studied at the level of cognitive processes in reinforcement learning methods.

Attractive and interesting robotic artifacts have been built able to convey emotions, although it is uncertain whether these experiments represent effective steps forward in understanding emotions. For example, the well known robot Kismet [4] is able to modify some of its external appearance like raising an eyebrow, grimace, and so on. during its interactions with an user. These simple external modifications are associates with emotions. Actually, Kismet has no real model of emotions, but merely uses a repertoire of rules defined in advance by the designer: it is the user that naively, interacting with the robot, ends up with the attribution of emotions to Kismet. On the other hand, it is the human tendency to anthropomorphize aspects of its environment. It is easy to see a pair of eyes and a mouth in a random shape, so it is at the same time easy to ascribe emotions and intentions to the actions of an agent.

In summary, an agent capable of transmitting emotions during jazz improvisation must have some effective computational models for generation and evocation of emotions.

6 EXPERIENCE

Finally, the more complex problem for consciousness is: how can a physical system like an agent able to improvise jazz to produce something similar to our subjective experience? During a jam session, the sound waves generated by the musical instruments strike our ears and we experience a sax solo accompanied by bass, drums and piano. At sunset, our retinas are struck by rays of light and we have the experience of a symphony of colors. We swallow molecules of various kinds and, therefore, we feel the taste of a delicious wine.

It is well known that Galileo Galilei suggested that smells, tastes, colors and sounds do not exist outside the body of a conscious subject (the *living animal*). Thus experience would be created by the subject in some unknown way.

A possible hypothesis concerns the separation between the domain of experience, namely, the subjective content, and the domain of objective physical events. The claim is that physical reality can be adequately described only by the quantitative point of view in a third person perspective while ignoring any qualitative aspects. After all, in a physics textbook there are many mathematical equations that describe a purely quantitative reality. There is room for quality content, feelings or emotions. Explaining these qualitative contents is the *hard* problem of consciousness [7].

Yet scholars as Strawson [37] questioned the validity of such a distinction as well as the degree of real understanding of the nature of the physical world.

Whether the mental world is a special construct generated by some feature of the nervous systems of mammals, is still an open question. It is fair to stress that there is neither empirical evidence nor theoretical arguments supporting such a view. In the lack of a better theory, we could also take into consideration the idea inspired by *externalism* [31] [32] according to which the physical world comprehends also those features that we usually attribute to the mental domain. A *physicalist* must be held that if something is real, and we assume consciousness is real, it has to be physical. Hence, in principle, a device can envisage it.

In the case of artificial agents for jazz improvisation, how is it possible to overcome the distinction between function and experience? Such a typical agent is made up by a set of interconnected modules, each operating in a certain way. How the operation of some or all of the interconnected modules should generate conscious experience? However, the same question could be transferred to the activity of neurons. Each neuron, taken alone, does not work differently from a software module or a chip. But it could remains a possibility: it is not the problem of the physical world, but of our theories of the physical world. Artificial agents are part of the same physical world that produce consciousness in human subjects, so they may exploit the same properties and characteristics that are relevant for conscious experience.

In this regard, Tononi [41] proposed a theory supported by results from neuroscience, according to which the degree of conscious experience is related to the amount of integrated information. According to this framework, the primary task of the brain is to integrate information and, noteworthy, this process is the same whether it takes place in humans or in artifacts like agents for jazz improvisation. According to this theory, conscious experience has two main characteristics. On the one side, conscious experience is differentiated because the potential set of different conscious states is huge. On the other side, conscious experience is integrated; in facts a conscious state is experienced as a single entity. Therefore, the substrate of conscious experience must be an integrated entity able to differentiate among a big set of different states and whose informational state is greater than the sum of the informational states of the component sub entities [1] [2].

According to this theory, Koch and Tononi [24] propose a potential new Turing test based on the integration of information: artificial systems should be able to mimic the human being not in language skills (as in the classic version of Turing test), but rather in the ability to integrate information from different sources.

Therefore, an artificial agent aware of its jazz improvisation should be able to integrate during time the information generated by its own played instrument, the instruments of its band as well as information from the body, i.e., the feedback from skin receptors, the receptors of the tendons and muscles and possibly from the vestibular apparatus. Furthermore, it should also be able also to integrate information related to emotions.

Some of the early studies based on suitable neural networks for music generation [40] are promising in the way to implement an information integration agent. However, we must emphasize the fact that the implementation of a true information integration system is a real technological challenge In fact, the typical engineering techniques for the building of an artifact is essentially based on the principle of *divide et impera*, that involves the design of a complex system through the decomposition of the system into easier smaller subsystems. Each subsystem then communicates with the others subsystems through well-defined interfaces so that the interaction between the subsystems happen in a very controlled way. Tononi's theory requires instead maximum interaction between the subsystems in order to allow an effective integration. Therefore, new techniques are required to design effective conscious agents.

Information integration theory raised heated debates in the scientific community. It could represent a first step towards a theoretically well-founded approach to machine consciousness. The idea of being able to find the *consciousness equations* which, like the Maxwell's equations in physics, are able to explain consciousness in living beings and in the artifacts, would be a kind of ultimate goal for scholars of consciousness.

7 CONCLUSIONS

The list of problems related to machine consciousness that have not been properly treated is long: the sensorimotor experience in improvisation, the sense of time in musical performance, the problem of the meaning of a musical phrase, the generation of musical mental images and so on. These are all issues of great importance for the creation of a conscious agent for jazz improvisation, although some of them may overlap in part with the arguments discussed above.

Although the classic AI achieved impressive results, and the program EMI by Cope is a great example, so far these issues have been addressed only partially.

In this article we have discussed the main issues to be addressed in order to design and build an artificial that can perform a jazz improvisation. The physicality, the ability to be located, to have emotions, to have some form of experience are all problems inherent in the problem of consciousness.

A new Turing test might be based on imitating the ability to distinguish a jazz improvisation produced by an artificial agent, maybe able to integrate information according to Tononi, than improvisation produced by an expert jazz musician.

As should be clear, this is a very broad subject that significantly extends the traditional the mind-brain problem.

Machine consciousness is, at the same time, a theoretical and technological challenge that forces to deal with old problems and new innovative approaches. It is possible, and hope that the artificial consciousness researchers push to re-examine many threads left hanging from the Artificial Intelligence and cognitive science. “Could consciousness be a theoretical time bomb, ticking away in the belly of AI? Who can say?” (Haugeland [18], p. 247).

REFERENCES

[1] D. Balduzzi and G. Tononi, 'Integrated information in discrete dynamical systems: Motivation and theoretical framework', *PLoS Computational Biology*, **4**, e1000091, (2008).

[2] D. Balduzzi and G. Tononi, 'Qualia: The geometry of integrated information', *PLoS Computational Biology,* **5**, e1000462, (2009).

[3] M. Boden, *The Creative Mind: Myths and Mechanisms - Second Edition,* Routledge, London, 2004.

[4] C. Breazeal, *Designing Sociable Robots,* MIT Press, Cambridge, MA, 2002.

[5] R. Brooks, *Cambrian Intelligence: The Early History of the New AI,* MIT Press, Cambridge, MA, 1999.

[6] Camurri, S. Hashimoto, M. Ricchetti, A. Ricci, K. Suzuki, R. Trocca and G. Volpe, 'EyesWeb: Toward Gesture and Affect Recognition in Interactive Dance and Music Systems', *Computer Music Journal*, **24**, 57 – 69, (2000).

[7] D. Chalmers. *The Conscious Mind: In Search of a Fundamental Theory,* Oxford University Press, Oxford, 1996.

[8] A. Chella and R. Manzotti (eds.), *Artificial Consciousness*, Imprint Academic, Exeter, UK, 2007.

[9] A. Chella and R. Manzotti, 'Machine Consciousness: A Manifesto for Robotics', *International Journal of Machine Consciousness*, **1**, 33 – 51, (2009).

[10] J. Coker, *Improvising Jazz*, Simon & Schuster, New York, NY, 1964.

[11] D. Cope, 'Computer Modeling of Musical Intelligence in EMI', *Computer Music Journal*, **16**, 69 – 83, 1992.

[12] D. Cope, *Virtual Music*. MIT Press, Cambridge, MA, 2001.

[13] D. Cope, *Computer Models of Musical Creativity*, MIT Press, Cambridge, MA, 2005.

[14] A. Damasio, *The Feeling of What Happens: Body and Emotion in the Making of Consciousness,* Houghton Mifflin Harcourt, 1999.

[15] A. Danto, 'The Transfiguration of Commonplace', *The Journal of Aesthetics and Art Criticism*, **33**, 139 – 148, (1974).

[16] J.-M. Fellous and M. A. Arbib, *Who Needs Emotions?: The Brain Meets the Robot*, Oxford University Press, Oxford, UK, 2005.

[17] A. Gabrielsson and P.N. Juslin, 'Emotional Expression in Music Performance: Between the Performer's Intention and the Listener's Experience', *Psychology of Music*, **24**, 68 – 91, (1996).

[18] J. Haugeland, *Artificial Intelligence: The Very Idea*, MIT Press, Bradford Books, Cambridge, MA, 1985.

[19] D. Hofstadter, 'Essay in the Style of Douglas Hofstadter', *AI Magazine*, Fall, 82 – 88, (2009).

[20] P.N. Johnson-Laird, 'Jazz Improvisation: A Theory at the Computational Level', in: *Representing Musical Structure*, P. Howell, R. West & I. Cross (eds.), Academic Press, London, 1991.

[21] P. N. Juslin & J. A. Sloboda (eds.), *Handbook of Music and Emotion - Theory, Research, Application,* Oxford University Press, Oxford, UK, 2010.

[22] P.N. Juslin & D. Västfjäll, 'Emotional responses to music: The need to consider underlying mechanisms', *Behavioral and Brain Sciences*, **31**, 559 – 621, (2008).

[23] K. Koch, 'Final Report of the Workshop *Can a Machine be Conscious'*, The Banbury Center, Cold Spring Harbor Laboratory, http://theswartzfoundation.com/abstracts/2001_summary.asp (last access 12/09/2011).

[24] K. Koch and G. Tononi, 'Can Machines Be Conscious?', *IEEE Spectrum*, June, 47 – 51, (2008).

[25] A. Koestler, *The Act of Creation*, London, Hutchinson, 1964.

[26] J. W. Krueger, 'Enacting Musical Experience', *Journal of Consciousness Studies*, **16**, 98 – 123, (2009).

[27] J. Manzolli and P.F.M.J. Verschure, 'Roboser: A Real-World Composition System', *Computer Music Journal*, **29**, 55 – 74, (2005).

[28] F. Pachet, 'Beyond the Cybernetic Jam Fantasy: The Continuator', *IEEE Computer Graphics and Applications,* January/February, 2 – 6, (2004).

[29] J. Pressing, 'Improvisation: Methods and Models', in: *Generative Processes in Music: The Psychology of Performance, Improvisation, and Composition*, J. Sloboda (ed.), Oxford University Press, Oxford, UK, 1988.

[30] P. Robbins & M. Aydede (eds.), *The Cambridge Handbook of Situated Cognition*, Cambridge, Cambridge University Press, 2009.

[31] T. Rockwell, *Neither Brain nor Ghost*, MIT Press, Cambridge, MA, 2005.

[32] M. Rowlands, *Externalism – Putting Mind and World Back Together Again*, McGill-Queen's University Press, Montreal and Kingston, 2003.

[33] G. Schuller, 'Forewords', in: *Improvising Jazz*, J. Coker, Simon & Schuster, New York, NY, 1964.

[34] J. R. Searle, 'Minds, brains, and programs', *Behavioral and Brain Sciences*, **3**, 417 – 457, (1980).

[35] A. Seth, 'The Strength of Weak Artificial Consciousness', *International Journal of Machine Consciousness*, **1**, 71 – 82, (2009).

[36] R. J. Sternberg (eds.), *Handbook of Creativity*, Cambridge, Cambridge University Press, 1999.

[37] G. Strawson, 'Does physicalism entail panpsychism?', *Journal of Consciousness Studies*, **13**, 3 – 31, (2006).

[38] J. Sundberg and R.T. Verrillo, 'Somatosensory Feedback in Musical Performance', (Editorial), *Music Perception: An Interdisciplinary Journal*, **9**, 277 – 280, (1992).

[39] N.P. McAngus Todd, 'Vestibular Feedback in Musical Performance: Response to «Somatosensory Feedback in Musical Performance»', *Music Perception: An Interdisciplinary Journal*, **10**, 379 – 382, (1993).

[40] P.M. Todd & D. Gareth Loy (eds.), *Music and Connectionism*, MIT Press, Cambridge, MA, 1991.

[41] G. Tononi, 'An Information Integration Theory of Consciousness', *BMC Neuroscience*, **5**, (2004).

Taking Turing Seriously (But Not Literally)

William York[1] and **Jerry Swan**[2]

Abstract. Results from present-day instantiations of the Turing test, most notably the annual Loebner Prize competition, have fueled the perception that the test is on the verge of being passed. With this perception comes the misleading implication that computers are nearing human-level intelligence. As currently instantiated, the test encourages an adversarial relationship between contestant and judge. We suggest that the underlying purpose of Turing's test would be better served if the prevailing focus on trickery and deception were replaced by an emphasis on transparency and collaborative interaction. We discuss particular examples from the family of Fluid Concepts architectures, primarily Copycat and Metacat, showing how a modified version of the Turing test (described here as a "modified Feigenbaum test") has served as a useful means for evaluating cognitive-modeling research and how it can suggest future directions for such work.

1 INTRODUCTION; THE TURING TEST IN LETTER AND SPIRIT

> The method of "postulating" what we want has many advantages; they are the same as the advantages of theft over honest toil. – Bertrand Russell, *Introduction to Mathematical Philosophy*

> **Interrogator:** Yet Christmas is a Winter's day, and I do not think Mr. Pickwick would mind the comparison.
> **Respondent:** LOL – *Pace* Alan Turing, "Computing Machinery and Intelligence"

If Alan Turing were alive today, what would he think about the Turing test? Would he still consider his imitation game to be an effective means of gauging machine intelligence, given what we now know about the Eliza effect, chatbots, and the increasingly vacuous nature of interpersonal communication in the age of texting and instant messaging?

One can only speculate, but we suspect he would find current instantiations of his eponymous test, most notably the annual Loebner Prize competition, to be disappointingly literal-minded. Before going further, it will help to recall Turing's famous prediction about the test from 1950:

> I believe that in about fifty years' time it will be possible, to programme computers, with a storage capacity of about 10^9, to make them play the imitation game so well that an average interrogator will not have more than 70 per cent chance of making the right identification after five minutes of questioning ([22], p. 442).

The Loebner Prize competition adheres closely to the outward form—or *letter*—of this imitation game, right down to the five-minute interaction period and (at least for the ultimate Grand Prize) the 70-percent threshold.[3] However, it is questionable how faithful the competition is to the underlying purpose—or *spirit*—of the game, which is, after all, to assess whether a given program or artifact should be deemed intelligent, at least relative to human beings.[4]

More generally, we might say that the broader purpose of the test is to assess progress in AI, or at least that subset of AI that is concerned with modeling human intelligence. Alas, this purpose gets obscured when the emphasis turns from pursuing this long-term goal to simply "beating the test." Perhaps this shift in emphasis is an inevitable consequence of using a behavioral test: "If we don't want that," one might argue, "then let us have another test." Indeed, suggestions have been offered for modifying the Turing test (cf. [6], [7], [3]), but we still see value in the basic idea behind the test—that of using observable "behavior" to infer underlying mechanisms and processes.

1.1 Priorities and payoffs

The letter–spirit distinction comes down to a question of research priorities, of short-term versus long-term payoffs. In the short term, the emphasis on beating the test has brought programs close to "passing the Turing test" in its Loebner Prize instantiation. Brian Christian, who participated in the 2009 competition as a confederate (i.e., one of the humans the contestant programs are judged against) and described the experience in his recent book *The Most Human Human*, admitted to a sense of urgency upon learning that "at the 2008 contest..., the top program came up shy of [passing] by just a single vote" ([1], p. 4). Yet in delving deeper into the subject, Christian realized the superficiality—the (near) triumph of "pure technique"—that was responsible for much of this success.

But it is not clear that the Loebner Prize has steered researchers toward any sizable long-term payoffs in understanding human intelligence. After witnessing the first Loebner Prize competition in 1991, Stuart Shieber [20] concluded, "What is needed is not more work on solving the Turing Test, as promoted by Loebner, but more work on the basic issues involved in understanding intelligent behavior. The parlor games can be saved for later" (p. 77). This conclusion seems as valid today as it was two decades ago.

1.2 Communication, transparency, and the Turing test

The question, then, is whether we might better capture the spirit of Turing's test through other, less literal-minded means. Our answer is

[1] Indiana University, United States, email: wwyork@indiana.edu
[2] University of Stirling, Scotland, email: jsw@cs.stir.ac.uk

[3] Of course, the year 2000 came and went without this prediction coming to pass, but that is not at issue here.
[4] See [5] for more discussion of the distinction between human-like intelligence versus other forms of intelligence in relation to the Turing test.

not only that we can, but that we must. The alternative is to risk trivializing the test by equating "intelligence" with the ability to mimic the sort of context-neutral conversation that has increasingly come to pass for "communication." Christian points out that "the Turing test is, at bottom, about the act of communication" ([1], p. 13). Yet given the two-way nature of communication, it can be hard to disentangle progress in one area (AI) from deterioration in others. As Jaron Lanier recently put it,

> You can't tell if a machine has gotten smarter or if you've just lowered your standards of intelligence to such a degree that the machine seems smart. If you can have a conversation with a simulated person presented by an AI program, can you tell how far you've let your sense of personhood degrade in order to make the illusion work for you? ([13], p. 32).

In short, the Turing test's reliance on purely verbal behavior renders it susceptible to tricks and illusions that its creator could not have reasonably anticipated. Methodologies such as statistical machine learning, while valuable as computational and engineering *tools*, are nonetheless better suited to *modeling* human banality than they are human intelligence. Additionally, the test, as currently instantiated, encourages an adversarial approach between contestant and judge that does as much to obscure and inflate progress in AI as it does to provide an accurate measuring stick. It is our contention that a test that better meets Turing's original intent should instead be driven by the joint aims of collaboration and transparency.

2 INTELLIGENCE, TRICKERY, AND THE LOEBNER PRIZE

Does deception presuppose intelligence on the part of the deceiver? In proposing his imitation game, Turing wagered—at least implicitly—that the two were inseparable. Surely, a certain amount of cunning and intelligence are required on the part of humans who excel at deceiving others. The flip side of the coin is that a degree of gullibility is required on the part of the person(s) being deceived.

Things get more complicated when the deception is "perpetrated" by a technological artifact as opposed to a willfully deceptive human. To quote Shieber once again, "[I]t has been known since Weizenbaum's surprising experiences with ELIZA that a test based on fooling people is confoundingly simple to pass" (p. 72; cf. [24]). The gist of Weizenbaum's realization is that our interactions with computer programs often tell us less about the inner workings of the programs themselves than they do about our tendency to project meaning and intention onto artifacts, even when we *know* we should know better.

2.1 The parallel case of art forgery

For another perspective on the distinction between genuine accomplishment and mere trickery, let us consider the parallel case of art forgery. Is it possible to distinguish between a genuine artist and a mere faker? It is tempting to reply that in order to be a good faker—one good enough to fool the experts—one must necessarily be a good artist to begin with. But this sort of argument is too simplistic, as it equates artistry with technical skill and prowess, meanwhile ignoring originality, artistic vision, and other qualities that are essential to genuine artistry (cf. [14], [2]). In particular, the ability of a skilled art forger to create a series of works in the style of, say, Matisse does not necessarily imply insight into the underlying artistic or expressive *vision* of Matisse—the vision responsible for giving rise to those works in the first place. As philosopher Matthew Kieran succinctly puts it, "There is all the difference in the world between a painting that genuinely reveals qualities of mind to us and one which blindly apes their outward show" ([11], p. 21).

Russell's famous quote about postulation equating to theft helps us relate an AI methodology to the artistry–forgery distinction. Russell's statement can be paraphrased as follows: merely saying that there exists a function (e.g., `sqrt()`) with some property (e.g., `sqrt(x)*sqrt(x)=x for all x >= 0`) does not tell us very much about how to *generate* the actual `sqrt()` function. Similarly, the ability to reproduce a small number of values of `x` that meet this specification does not imply insight into the underlying *mechanisms* involved, relative to which the existence of these specific values is essentially a side effect. A key issue here is the small number of values: Since contemporary versions of the Turing test are generally highly time-constrained, it is even more imperative that the test involve a deep probe into the possible behaviors of the respondent.

2.2 Thematic variability in art and in computation

Many of the Loebner Prize entrants (e.g., [23]) have adopted the methodologies of corpus linguistics and machine learning, so let us reframe the issue of thematic variability in these terms. We might abstractly consider the statistical machine-learning approach to the Turing test as being concerned with the induction of a generative grammar. In short, the ability to induce an algorithm that reproduces some themed collection of original works does not in itself imply that any underlying sensibilities that *motivated* those works can be effectively approximated by that algorithm.

One way of measuring the "work capacity" of an algorithm is to employ the Kolmogorov complexity measure [21], which is essentially the size of the shortest possible functionally identical algorithm. In the induction case, algorithms with the lowest Kolmogorov complexity will tend to be those that exhibit very little variability—in the limiting case, generating only instances from the original collection. This would be analogous to a forger who could only produce exact copies of another artist's works, rather than works "in the style of" said artist—the latter being the stock-in-trade of infamous art forgers Han van Meegeren [25] and Elmyr de Hory [10].

In contrast, programs from the family of Fluid Concepts architectures (see 4.1 below) possess relational and generative models that are domain-specific. For example, the Letter Spirit architecture [19] is specifically concerned with exploring the thematic variability of a given font style. Given Letter Spirit's (relatively) sophisticated representation of the "basis elements" and "recombination mechanisms" of form, it might reasonably be expected to have a high Kolmogorov complexity. The thematic variations generated by Letter Spirit are therefore not easily approximated by domain-agnostic data-mining approaches.

2.3 Depth, shallowness, and the Turing test

The artistry–forgery distinction is useful insofar as it offers another perspective on the issue of depth versus shallowness—an issue that is crucial in any analysis of the Turing test. Just as the skilled art forger is adept at using trickery to simulate "authenticity"—for example, by artificially aging a painting through various techniques such as baking or varnishing ([10], [25])—analogous forms of trickery tend to find their way into the Loebner Prize competition: timely pop-culture references, intentional typos and misspellings, strategic changes of

subject, and so on (cf. [20], [1]). Yet these surface-level tricks have as much to do with the genuine modeling of intelligence as coating the surface of a painting with antique varnish has to do with *bona fide* artistry. Much like the art forger's relationship with the art world, the relationship between contestant programs and judges in the Loebner Prize is essentially adversarial, not collaborative. The adversarial nature of these contestant–judge interactions, we feel, is a driving force in the divergence of the Turing test, in its current instantiations, from the spirit in which it was originally conceived.

3 SOME VARIATIONS ON THE TURING TEST

The idea of proposing modifications to the Turing test is not a new one. In this section, we look at such proposals—Stevan Harnad's "Total Turing Test" (and the accompanying hierarchy of Turing tests he outlines) and Edward Feigenbaum's eponymous variation on the Turing test—before discussing how they relate to our own, described below as a "modified Feigenbaum test."

3.1 The Total Turing Test

Harnad ([6], [7]) has outlined a detailed hierarchy of possible Turing tests, with Turing's own version occupying the second of five rungs on this hypothetical ladder. Harnad refers to this as the T2, or "pen-pal," level, given the strict focus on verbal (i.e., written or typed) output. Directly below this level is the t1 test (where "t" stands for "toy," not "Turing"). Harnad observed, a decade ago, that "all of the actual mind-modelling research efforts to date are still only at the t1 level, and will continue to be so for the foreseeable future: Cognitive Science has not even entered the TT hierarchy yet" ([7], §9). This is still the case today.

Just as the t1 test draws on "subtotal fragments" of T2, T2 stands in a similar relation to T3, the Total Turing Test. This test requires not just pen-pal behavior, but robotic (i.e., embodied) behavior as well. A machine that passed the Total Turing Test would be functionally (though not microscopically) indistinguishable from a human being.[5]

Clearly, there are fewer degrees of freedom—and hence less room for deception—as we climb the rungs on Harnad's ladder, particularly from T2 to T3. However, given the current state of the art, the T3 can only be considered an extremely distant goal at this point. It may be that the T2, or pen-pal, test could only be convincingly "passed"—over an arbitrarily long period of time, as Harnad stipulates, and not just the five-minute period suggested by Turing and adhered to in the Loebner Prize competition—by a system that could move around and interact with other people and things in the real world as we do. It may even be that certain phenomena that are still being modeled and tested at the t1 level—even seemingly abstract and purely "cognitive" ones such as analogy-making and categorization—are ultimately grounded in embodiment and sensorimotor capacities as well (cf. [12]), which would imply fundamental limitations for much current research. Unfortunately, such questions must be set aside for the time being, as they are beyond the scope of this paper.

3.2 The Feigenbaum test

The Feigenbaum test [3] was proposed in order test the quality of reasoning in specialized domains—primarily scientific or otherwise technical domains such as astrophysics, computer science, and medicine. The confederate in the Feigenbaum test is not merely an ordinary human being, but an "elite scientist" and member of the U.S. National Academy of Sciences. The judge, who is also an Academy member and an expert in the domain in question, interacts with the confederate and the contestant (i.e., the program). Feigenbaum elaborates, "The judge poses problems, asks questions, asks for explanations, theories, and so on—as one might do with a colleague" ([3], p. 36). No time period is stipulated, but as with the Turing test, "the challenge will be considered met if the computational intelligence 'wins' one out of three disciplinary judging contests, that is, one of the three judges is not able to choose reliably between human and computer performer" (ibid.).

3.3 A modified Feigenbaum test

Feigenbaum's emphasis on knowledge-intensive technical domains is in keeping with his longtime work in the area of expert systems. This aspect of his test is incidental, even irrelevant, to our purposes. In fact, we go one step further with our "modified Feigenbaum test" and remove the need for an additional contestant beyond the program. Rather, the judge "interacts" directly with the program for an arbitrarily long period of time and evaluates the program's behavior directly—and qualitatively—on the basis of this interaction. (No illusion is made about the program passing for human, which would be premature and naive in any case.)

What *is* relevant about the Feigenbaum test for our purposes is its emphasis on focused, sustained interaction between judge and program within a suitably subtle domain. Our modified Feigenbaum test stresses a similar type of interaction, though the domain—while still constrained—is far less specialized or knowledge-intensive than, say, astrophysics or medicine. In fact, the domain we discuss below—letter-string analogies—was originally chosen as an arena for modeling cognition because of its balance of generality and tractability [9]. In other words, the cognitive processes involved in thinking and otherwise "operating" within the domain are intended to be more or less general and domain-independent. At the same time, the restriction of the domain, in terms of the entities and relationships that make it up, is meant to ensure tractability and plausibility—in contrast to dealing (or pretending to deal) with complex real-world knowledge of a sort that can scarcely be attributed to a computer program (e.g., knowledge of medicine, the solar system, etc.).

In the following section, we argue on behalf of this approach and show how research carried out under this ongoing program represents an example of how one can take the idea of Turing's test seriously without taking its specifications literally.

4 TAKING TURING SERIOUSLY: AN ALTERNATIVE APPROACH

In an essay entitled "On the Seeming Paradox of Mechanizing Creativity," Hofstadter [8] relates Myhill's [17] three classes of mathematical logic to categories of behavior. The most inclusive category, the *productive*, is the one that is of central interest to us here. While no finite collection of rules suffices to generate the members of a productive set P (and no $x \notin P$), a more expansive and/or sophisticated set of generative rules (i.e., creative processes) can approximate P with unbounded accuracy.

In order to emphasize the role of such "unbounded creativity" in the evaluation of intelligence, we describe a modified Feigenbaum test restricted to the microdomain of letter-string analogies. An example of such a problem is, "If **abc** changes to **abd**, how would you change **pxqxrx** in 'the same way'?" (or simply **abc** $\rightarrow$ **abd**; **pxqxrx**

[5] The T4 and T5 levels, which make even greater demands, are not relevant for our purposes.

→ **???**). Problems in this domain have been the subject of extensive study [9], resulting in the creation of the well-known Copycat model [16] and its successor, Metacat [15]. Before describing this test, however, we briefly discuss these programs' architectures in general terms.

4.1 Copycat, Metacat, and Fluid Concepts architectures

Copycat's architecture consists of three main components, all of which are common to the more general Fluid Concepts architectural scheme. These components are the Workspace, which is essentially roughly the program's working memory; the Slipnet, a conceptual network with variably weighted links between concepts (essentially a long-term memory); and the Coderack, home to a variety of agent-like codelets, which perform specific tasks in (simulated) parallel, without the guidance of an executive controller. For example, given the problem **abc** → **abd**; **iijjkk** → **???**, these tasks would range from identifying groups (e.g., the **jj** in **iijjkk**) to proposing bridges between items in different letter-strings (e.g., the **b** in **abc** and the **jj** in **iijjkk**) to proposing rules to describe the change in the initial pair of strings (i.e., the change from **abc** to **abd**).[6]

Building on Copycat, Metacat incorporates some additional components that are not present in its predecessor's architecture, most notably the Episodic Memory and the Temporal Trace. As the program's name suggests, the emphasis in Metacat is on *metacognition*, which can broadly be defined as the process of monitoring, or thinking about, one's own thought processes. What this means for Metacat is an ability to monitor, via the Temporal Trace, events that take place en route to answering a given letter-string problem, such as detecting a "snag" (e.g., trying to find the successor to **z**, which leads to a snag because the alphabet does not "circle around" in this domain) or noticing a key idea. Metacat also keeps track of its answers to previous problems, as well as its responses on previous runs of the same problem, both via the Episodic Memory. As a result, it is able to be "reminded" of previous problems (and answers) based on the problem at hand. Finally, it is able to compare and contrast two answers at the user's prompting (see Section 4.3 below).

Philosophically speaking, Fluid Concepts architectures are predicated upon the conviction that it is possible to "know everything about" the entities and relationships in a given microdomain. In other words, there is no propositional fact about domain entities and processes (or the effect of the latter on the on the former) that is not in principle accessible to inspection or introspection. In Copycat, the domain entities range from permanent "atomic" elements (primarily, the 26 letters of the alphabet) to temporary, composite ones, such as the letter strings that make up a given problem (**abc**, **iijjkk**, **pxqxrx**, etc.); the groups within letter strings that are perceived during the course of a run (e.g., the **ii**, **jj**, and **kk** in **iijjkk**); and the bonds that are formed between such groups. The relationships include concepts such as *same*, *opposite*, *successor*, *predecessor*, and so on.

A key aspect of the Fluid Concepts architecture is that it affords an exploration the space of instantiations of those entities and relationships in a (largely) non-stochastic fashion—that is, in a manner that is predominately directed by the nature of the relationships themselves. In contrast, the contextual pressures that give rise to some subtle yet low frequency solutions are unlikely to have a referent within a statistical machine-learning model built from a corpus of Copycat answers, since outliers are not readily captured by gross mechanisms such as sequences of transition probabilities.

[6] See [16] for an in-depth discussion of codelet types and functions in Copycat.

4.2 An example from the Copycat microdomain

To many observers, a letter-string analogy problem such as the aforementioned **abc** → **abd**; **iijjkk** → **???** might appear trivial on first glance.[7] Yet upon closer inspection, one can come to appreciate the surprising subtleties involved in making sense of even a relatively basic problem like this one. Consider the following (non-exhaustive) list of potential answers to the above problem:

- **iijjll** – To arrive at this seemingly basic answer requires at least three non-trivial insights: (1) seeing **iijjkk** as a sequence of three sameness groups—**ii**, **jj**, and **kk**—not as a sequence of individual letters; (2) seeing the group **kk** as playing the same role in **iijjkk** that the letter **c** does in **abc**; and (3) seeing the change from **c** to **d** in terms of *successorship* and not merely as a change from the *letter* **c** to the *letter* **d**. The latter point may seem trivial, but it is not a given, and as we will see, there are other possible interpretations.
- **iijjkl** – This uninspiring answer results from simply changing the letter category of the rightmost *letter* in **iijjkk** to its successor, as opposed to the letter category of the rightmost *group*.
- **iijjkd** – This answer results from the literal-minded strategy of simply changing the last letter in the string to **d**, all the while ignoring the other relationships among the various groups and letter categories.
- **iijjdd** – This semi-literal, semi-abstract answer falls somewhere in between **iijjll** and **iijjkl**. On the one hand, it reflects a failure to perceive the change from **c** to **d** in the initial string in terms of *successorship*, instead treating it as a mere replacement of the *letter* **c** with the *letter* **d**. On the other hand, it does signal a recognition that the concept *group* is important, as it at least involves carrying out the change from **k** to **d** in the target string over to *both* **k**s and not just the rightmost one. This answer has a "humorous" quality to it, unlike **iijjkl** or **iijjkd**, due to its mixture of insight and confusion.

This incomplete catalog of answers hints at the range of issues that can arise in examining a single problem in the letter-string analogy domain. Copycat itself is able to come up with all of the aforementioned answers (along with a few others), as illustrated in Table 1, which reveals **iijjll** to be the program's "preferred choice" according to the two available measures. These measures are (1) the relative frequency with which each answer is given and (2) the average "final temperature" associated with each answer. Roughly speaking, the temperature—which can range from 0 to 100—indicates the program's moment-to-moment "happiness" with its perception of the problem during a run, with a lower temperature corresponding to a more positive evaluation

4.3 The modified Feigenbaum test: from Copycat to Metacat

One limitation of Copycat is its inability to "say" anything about the answers it gives beyond what appears in its Workspace during the

[7] Such problems may seem to bear a strong resemblance to the kinds of problems one might find on an IQ test. However, an important difference worth noting is that the problems in the Copycat domain are not conceived of as having "correct" or "incorrect" answers (though in many cases there are clearly "better" and "worse" ones). Rather, the answers are open to discussion, and the existence of subtle differences between the various answers to a given problem is an important aspect of the microdomain.

Table 1. Copycat's performance over 1000 runs on the problem **abc** → **abd**; **iijjkk** → **???**. Adapted from [16].

Answer	Frequency	Average Final Temperature
iijjll	810	27
iijjkl	165	47
iijjdd	9	32
iikkll	9	46
iijkll	3	43
iijjkd	3	65
ijkkll	1	43

course of a run. While aggregate statistics such as those illustrated in Table 1 can offer some insight into its performance, the program is not amenable to genuine Feigenbaum-testing, primarily because it doesn't have the capacity to summarize its viewpoint. To the extent that it *can* be Feigenbaum-tested, it can only do so in response to what might termed first-order questions (e.g., **abc** → **abd**; **iijjkk** → **???**). It cannot answer second-order questions (i.e., questions *about* questions), let alone questions about its *answers* to questions about questions.

In contrast, Metacat allows us to ask increasingly sophisticated questions of it, and thus can be said to allow for the sort of modified Feigenbaum-testing described in Section 3.3. One can "interact" with the program in a variety of ways: by posing new problems; by inputting an answer to a problem and running the program in "justify mode," asking it to evaluate and make sense of the answer; and by having it compare two answers to one another (as in the above examples). In doing the latter, the program summarizes its "viewpoint" with one of a set of canned (but non-arbitrary) English descriptions. For example, the preferred answer might be "based on a richer set of ideas," "more abstract," or "more coherent."

The program also attempts to "explain" how the two answers are similar to each other and how they differ. For example, consider the program's summary of the comparison between **iijjll** and **iijjdd** in response to the aforementioned problem:

> The only essential difference between the answer **iijjdd** and the answer **iijjll** to the problem **abc** → **abd**; **iijjkk** → **???** is that the change from **abc** to **abd** is viewed in a more literal way for the answer **iijjdd** than it is in the case of **iijjll**. Both answers rely on seeing two strings (**abc** and **iijjkk** in both cases) as groups of the same type going in the same direction. All in all, I'd say **iijjll** is the better answer, since it involves seeing the change from **abc** to **abd** in a more abstract way.

It should be emphasized that the specific form of the verbal output is extremely unsophisticated relative to the capabilities of the underlying architecture, indicating that it is possible to exhibit depth of insight while treating text generation as essentially a side-effect. This contrasts sharply with contemporary approaches to the Turing test.

For the sake of contrast, here is the program's comparison between the answers **iijjll** and **abd**, which illustrates some of the program's limitations in clumsily (and, of course, unintentionally) humorous fashion:

> The only essential difference between the answer **abd** and the answer **iijjll** to the problem **abc** → **abd**; **iijjkk** → **???** is that the change from **abc** to **abd** is viewed in a completely different way for the answer **abd** than it is in the case of **iijjll**. Both answers rely on seeing two strings (**abc** and **iijjkk** in both cases) as groups of the same type going in the same direction. All in all, I'd say **abd** is really terrible and **iijjll** is very good.

Apart from the thin veneer of human agency that results from Metacat's text generation, the program's accomplishments—and just as importantly, its *failures*—become transparent through interaction.

4.4 Looking ahead

In order for it to actually pass an "unrestricted modified Feigenbaum test" in the letter-string analogy domain, what other questions might we conceivably require Metacat to answer? Here are some suggestions:

1. Problems that involve more holistic processing of letter strings. There are certain letter strings that humans seem to have little trouble processing, but that are beyond Metacat's grasp—for example, the string **oooaaoobbooooccoo** in the problem **abc** → **abd**; **oooaaoobbooooccoo** → **???**. How are we so effortlessly able to "tune out" the **o**'s in **oooaaoobbooooccoo**? What would it take for a Metacat-style program to be able to do likewise?
2. Meta-level questions about sequences of answers. For example, "How is the relationship between answer A and answer B different from that between C and D?" Such questions could be answered using the declarative information that Metacat already has; all that would seem to be required is the ability to pose the question.
3. Questions pertaining to concepts about analogy-making in general, such as *mapping*, *role*, *theme*, *slippage*, *pressure*, *pattern*, and *concept*. Metacat deals implicitly with all of these ideas, but it doesn't have explicit knowledge or understanding of them.
4. An ability to characterize problems in terms of "the issues they are about," with the ultimate goal of having a program that is able to create new problems of its own—which would certainly lead to a richer, more interesting exchange between the program and the human interacting with it. Some work in this area was done in the Phaeaco Fluid Concepts architecture [4], but the issue requires further investigation.
5. Questions of the form, "Why is answer *A* more humorous (or stranger, or more elegant, etc.) than answer *B*?" Metacat has implicit notions, however primitive, of concepts such as *succinctness*, *coherence*, and *abstractness*, which figure into its answer comparisons. These notions pertain to aesthetic judgment insofar as we tend to find things that are succinct, coherent, and reasonably abstract to be more pleasing than things that are prolix, incoherent, and either overly literal or overly abstract. Judgments involving humor often take into account such factors, too, among many others. Metacat's ability—however rudimentary—to employ criteria such as abstractness and coherence in its answer evaluations could be seen as an early step toward understanding how these kinds of qualitative judgments might emerge from simpler processes. On the other hand, for adjectives such as "humorous," which presuppose the possession of emotional or affective states, it is not at all clear what additional mechanisms might be required, though some elementary possibilities are outlined in [18].
6. A rudimentary sense of the "personality traits" associated with certain patterns of answers. In other words, just as Metacat is able compare two answers with one another, a meta-Metacat might be able to compare two *sets* of answers—and, correspondingly, two *answerers*—with one another. For example, a series of literal-minded or short-sighted answers might yield a perception of the answerer as being dense, while a series of sharp, insightful an-

swers punctuated by the occasional obvious clunker might yield a picture of an eccentric smart-aleck.

Ultimately, however, the particulars of Copycat, Metacat, and the letter-string analogy domain are not so important in and of themselves. The programs merely serve as an example of a kind of approach to modeling cognitive phenomena, just as the domain itself serves as a controlled arena for carrying out such modeling.

To meet the genuine intent of the Turing test, we must be able to partake in the sort of arbitrarily detailed and subtle discourse described above in any domain. As the forgoing list shows, however, there is much that remains to be done, even—to stick with our example—within the tiny domain in which Copycat and Metacat operate. It is unclear how far a disembodied computer program, even an advanced successor to these two models, can go toward modeling socially and/or culturally grounded phenomena such as personality, humor, and aesthetic judgment, to name a few of the more obvious challenges involved in achieving the kind of discourse that our "test" ultimately calls for. At the same time, it is unlikely that such discourse lies remotely within the capabilities of any of the current generation of Loebner Prize contenders, nor does it even seem to be a goal of such contenders.

5 CONCLUSION

We have argued that the Turing test would more profitably be considered as a sequence of modified Feigenbaum tests, in which the questioner and respondent are to collaborate in an attempt to extract maximum subtlety from a succession of arbitrarily detailed domains. In addition, we have explored a parallel between the "domain-agnostic" approach of statistical machine learning and that of artistic forgery, in turn arguing that by requesting successive variations on an original theme, a critic may successfully distinguish mere surface-level imitations from those that arise via the meta-mechanisms constitutive of genuine creativity and intelligence. From the perspective we have argued for, Metacat and the letter-string-analogy domain can be viewed as a kind of Drosophila for the Turing test, with the search for missing mechanisms directly motivated by the specific types of questions we might conceivably ask of the program.

ACKNOWLEDGEMENTS

We would like to thank Vincent Müller and Aladdin Ayesh for their hard work in organizing this symposium, along with the anonymous referees who reviewed and commented on the paper. We would also like to acknowledge the generous support of Indiana University's Center for Research on Concepts and Cognition.

REFERENCES

[1] B. Christian, *The Most Human Human*, Doubleday, New York, 2011.
[2] D. Dutton, 'Artistic crimes', *British Journal of Aesthetics*, **19**, 302–314, (1979).
[3] E. A. Feigenbaum, 'Some challenges and grand challenges for computational intelligence', *Journal of the ACM*, **50(1)**, 32–40, (2003).
[4] H. Foundalis. Phaeaco: A cognitive architecture inspired by bongard's problems. Doctoral dissertation, Indiana Univ., Bloomington, 2006.
[5] R. French, 'Subcognition and the limits of the Turing test', *Mind*, **99**, 53–65, (1990).
[6] S. Harnad, 'The Turing test is not a trick: Turing indistinguishability is a scientific criterion', *SIGART Bulletin*, **3(4)**, 9–10, (1992).
[7] S. Harnad, 'Minds, machines and Turing: the indistinguishability of indistinguishables', *Journal of Logic, Language, and Information*, **9(4)**, 425–445, (2000).
[8] D. R. Hofstadter, *Metamagical Themas: Questing for the Essence of Mind and Pattern*, Basic Books, New York, 1986.
[9] D. R. Hofstadter, *Fluid Concepts and Creative Analogies*, Basic Books, New York, 1995.
[10] C. Irving, *Fake! The story of Elmyr de Hory, the greatest art forger of our time*, McGraw-Hill, New York, 1969.
[11] M. Kieran, *Revealing Art*, Routledge, London, 2005.
[12] B. Kokinov, V. Feldman, and I. Vankov, 'Is analogical mapping embodied?', in *New Frontiers in Analogy Research*, eds., B. Kokinov, K. Holyoak, and D. Gentner, New Bulgarian Univ. Press, Sofia, Bulgaria, (2009).
[13] J. Lanier, *You Are Not a Gadget*, Alfred A. Knopf, New York, 2010.
[14] A. Lessing, 'What is wrong with a forgery?', *Journal of Aesthetics and Art Criticism*, **23(4)**, 461–471, (1979).
[15] J. Marshall. Metacat: A self-watching cognitive architecture for analogy-making and high-level perception. Doctoral dissertation, Indiana Univ., Bloomington, 1999.
[16] M. Mitchell, *Analogy-Making as Perception: A Computer Model*, MIT Press, Cambridge, Mass., 1993.
[17] J. Myhill, 'Some philosophical implications of mathematical logic', *Review of Metaphysics*, **6**, 165–198, (1952).
[18] R. Picard, *Affective Computing*, MIT Press, Cambridge, Mass., 1997.
[19] J. Rehling. Letter spirit (part two): Modeling creativity in a visual domain. Doctoral dissertation, Indiana Univ., Bloomington, 2001.
[20] S. Shieber, 'Lessons from a restricted Turing test', *Communications of the ACM*, **37(6)**, 70–78, (1994).
[21] R.J. Solomonoff, 'A formal theory of inductive inference, pt. 1', *Information and Control*, **7(1)**, 1–22, (1964).
[22] A. Turing, 'Computing machinery and intelligence', *Mind*, **59**, 433–460, (1950).
[23] R. Wallace, 'The anatomy of A.L.I.C.E.', in *Parsing the Turing Test*, eds., R. Epstein, G. Roberts, and G. Beber, 1–57, Spring, Heidelberg, (2009).
[24] J. Weizenbaum, *Computer Power and Human Reason*, Freeman, San Francisco, 1976.
[25] H. Werness, 'Han van Meegeren fecit', in *The Forger's Art*, ed., D. Dutton, 1–57, Univ. of California Press, Berkeley, (1983).

Laws of Form and the Force of Function. Variations on the Turing Test

Hajo Greif[1]

Abstract. This paper commences from the critical observation that the Turing Test (TT) might not be best read as providing a definition or a genuine test of intelligence by proxy of a simulation of conversational behaviour. Firstly, the idea of a machine producing likenesses of this kind served a different purpose in Turing, namely providing a demonstrative simulation to elucidate the force and scope of his computational method, whose primary theoretical import lies within the realm of mathematics rather than cognitive modelling. Secondly, it is argued that a certain bias in Turing's computational reasoning towards formalism and methodological individualism contributed to systematically unwarranted interpretations of the role of the TT as a simulation of cognitive processes. On the basis of the conceptual distinction in biology between structural homology vs. functional analogy, a view towards alternate versions of the TT is presented that could function as investigative simulations into the emergence of communicative patterns oriented towards shared goals. Unlike the original TT, the purpose of these alternate versions would be co-ordinative rather than deceptive. On this level, genuine functional analogies between human and machine behaviour could arise in quasi-evolutionary fashion.

1 A Turing Test of What?

While the basic character of the Turing Test (henceforth TT) as a simulation of human conversational behaviour remains largely unquestioned in the sprawling debates it has triggered, there are a number of diverging interpretations as to whether and to what extent it provides a definition, or part of a definition, of intelligence in general, or whether it amounts to the design of an experimental arrangement for assessing the possibility of machine intelligence in particular. It thus remains undecided what role, if any, there is for the TT to play in cognitive inquiries.

I will follow James H. Moor [13] and other authors [21, 2] in their analysis that, contrary to seemingly popular perception, the TT does neither provide a definition nor an empirical criterion of the named kind. Nor was it intended to do so. At least at one point in Alan M. Turing's, mostly rather informal, musings on machine intelligence, he explicitly dismisses the idea of a definition, and he attenuates the idea of an empirical criterion of machine intelligence:

> I don't really see that we need to agree on a definition [of thinking] at all. The important thing is to try to draw a line between the properties of a brain, or of a man, that we want to discuss, and those that we don't. To take an extreme case, we are not interested in the fact that the brain has the consistency of cold porridge. We don't want to say 'This machine's quite hard, so it isn't a brain, and so it can't think.' I would like to suggest a particular kind of test that one might apply to a machine. You might call it a test to see whether the machine thinks, but it would be better to avoid begging the question, and say that the machines that pass are (let's say) 'Grade A' machines. [. . .] (Turing in a BBC radio broadcast of January 10th, 1952, quoted after [3, p. 494 f])

Turing then goes on to introducing a version of what has come to be known, perhaps a bit unfortunately, as the Turing Test, but was originally introduced as the "imitation game". In place of the articulation of definitions of intelligence or the establishment of robust empirical criteria for intelligence, we find much less ambitious, and arguably more playful, claims. One purpose of the test was to develop a thought-experimental, inductive approach to identifying those properties shared between the human brain and a machine which would actually matter to asking the question of whether men or machines alike can think: *What is the common ground human beings and machines would have to share in order to also share a set of cognitive traits?* It was not a matter of course in Turing's day that there could possibly be any such common ground, as cognition was mostly considered essentially tied to (biological or other) human nature.[2] In many respects, the TT was one very instructive and imaginative means of raising the question whether the physical constitution of different systems, whether cold-porrige-like or electric-circuitry-like, makes a principled difference between a system with and a system without cognitive abilities. Turing resorted to machine simulations of behaviours that would normally be considered expressions of human intelligence in order to demonstrate that the lines of demarcation between the human and the mechanical realm are less than stable.

The TT is however not sufficient as a means for *answering* the questions it first helped to raise, nor was it so intended. Turing's primary aim for the TT was one demonstration, among others, of the force and scope of what he introduced as the "computational method" (which will be briefly explained in section 2). Notably, the computational method has a systematically rooted bias towards, firstly, considering a system's logical form over its possible functions and towards, secondly, methodological individualism. I will use Turing's mathematical theory of morphogenesis and, respectively, the distinction between the concepts of structural homology and functional analogy in biology as the background for discussing the implications of this twofold bias (in section 3). On the basis of this discussion, a tentative reassessment of the potentials and limits of the

[1] University of Klagenfurt, Austria, email: hajo.greif@aau.at

[2] In [1, p. 168 f], Margaret Boden notices that the thought that machines could possibly think was not even a "heresy" up to the early 20th century, as that claim would have been all but incomprehensible.

TT as a simulation will be undertaken (in section 4): If there is a systematic investigative role to play in cognitive inquiries for modified variants of the TT, these would have to focus on possible functions to be shared between humans and machines, and they would have to focus on shared environments of interaction rather than individual behaviours.

2 The Paradigm of Computation

Whether intentionally or not, Turing's reasoning contributed to breaking the ground for the functionalist arguments that prevail in much of the contemporary philosophies of biology and mind: An analysis is possible of the operations present within a machine or an organism that systematically abstracts from their respective physical nature. An set of operations identical on a specified level of description can be accomplished in a variety of physical arrangements. Any inference from the observable behavioural traits of a machine simulating human communicative behaviour, as in the TT, to an identity of underlying structural features would appear unwarranted.

Turing's work was concerned with the possibilities of devising a common logical form of abstractly describing the operations in question. His various endeavours, from morphogenesis via (proto-) neuronal networks to the simulation of human conversational behaviour, can be subsumed under the objective of exploring what his "computational method" could achieve across a variety of empirical fields and under a variety of modalities. Simulations of conversational behaviours that had hitherto been considered an exclusively human domain constituted but one of these fields, investigated under one modality.

Turing's computational method is derived from his answer to a logico-mathematical problem, David Hilbert's "Entscheidungsproblem" (the decision problem) in predicate logic, as presented in [8]. This problem amounts to the question whether, within the confines of a logical calculus, there is an unequivocal, well-defined and finite, hence at least in principle executable, procedure for deciding on the truth of a proposition stated in that calculus. After Kurt Gödel's demonstration that neither the completeness nor the consistency of arithmetic could be proven or disproven within the confines of arithmetic proper [7], the question of deciding on the *truth* of arithmetical propositions from within that same axiomatic system had to be recast as a question of deciding on the internal *provability* of such propositions. The – negative – answer to this reformulated problem was given by Turing [18] (and, a little earlier, by a slightly different method, Alonzo Church). Turing's path towards that answer was based on Gödel's elegant solution to the former two problems, namely a translation into arithmetical forms of the logical operations required for deciding on the provability of that proposition within the system of arithmetical axioms. Accordingly, the method of further investigation was to examine the calculability of the arithmetical forms so generated.

To decide on the calculability of the problem in turn, Turing introduced the notion of computability. A mathematical problem is considered computable if the process of its solution can be broken down into a set of exact elementary instructions by which one will arrive at a determinate solution in a finite number of steps, and which could be accomplished, at least in principle, by human "computers".[3] Even complex problems should thus become reducible to a set of basic operations. The fulfilment of the required routines demands an ability to apply a set of rules and, arguably, some mental discipline, but these routines are not normally considered part of the most typical or complex properties of human thought – and can be mechanised, in a more direct, material sense, by an appropriately constructed and programmed machine. Hence, Turing's notion of "mechanical" was of a fairly abstract kind. It referred to a highly standardised and routinised method of solving mathematical problems, namely the computational method proper. This method could be equally applied by human, mechanical or digital "computers", or by any other system capable of following the required routines.

Given this description of computability, the primary aim of Turing's models of phenomena such as morphogenesis, the organisation of the nervous system or the simulation of human conversation lies in finding out whether, how and to what extent their specific structural or behavioural patterns can be formally described in computational terms – and thus within the realm of mathematics. A successful application of the computational method to the widest variety of phenomena would have implications on higher epistemological or arguably even metaphysical levels, but, being possible implications, these are not contained within the mathematical theory.

3 The Relevance of Form and Function

The design of Turing's computational method intuitively suggests, but does not entail, that the phenomena in question are chiefly considered in their, computationally modellable, *form*. Turing focuses on the formal patterns of organic growth, on the formal patterns of neuronal organisation and re-organisation in learning, and on the logical forms of human conversation. The possible or actual functions of these formally described patterns, in terms of the purposes they do or may serve, are not systematically considered. A second informal implication of Turing's computational approach lies in his focus on the behaviour of isolated, *individual* systems – hence not on the organism in its environment, but on the human brain as a device with input and output functions.[4] Such focus on self-contained, individual entities was arguably guided by a methodological presupposition informed by the systematic goals of Turing's research: The original topics of his inquiry were the properties of elementary recursive operations within a calculus. Hence, any empirical test for the force and scope of the computational method, that is, any test for what can be accomplished by means of such elementary recursive operations, would naturally but not necessarily commence in the same fashion.

In order to get a clearer view of this twofold bias, it might be worthwhile to take a closer look at the paradigm of Turing's computational method. That paradigm, in terms of elaboration, rigour and systematicity, is not to be found in his playful and informal imitation game approach to computer simulations of conversational behaviour. Instead, it is to be found in his mathematical theory of morphogenesis [20]. This inquiry was guided by Sir D'Arcy Thompson's, at its time, influential work *On Growth and Form* [17], and it was directed at identifying the basic chemical reactions involved in generating organic patterns, from an animal's growth to the grown animal's anatomy, from the dappledness or stripedness of furs to the arrangement of a sunflower's florets and the phyllotactic ordering of leaves on a plant's twigs. The generation of such patterns was modelled in rigorously formal-mathematical fashion. The resulting model was impartial to the actual biochemical realisation of pattern formation. It would only provide some cues as to what concrete reactants, termed "morphogens" by Turing, one should look out for.

[3] I am following B. Jack Copeland [4] here on his definition of computability, as he makes a considerable effort at spelling out what notion of computability Turing was using in [18]. He thus hopes to stem the often-lamented flood of loose and misguiding uses of that term in many areas of science.

[4] For this observation, see, for example, [9, p. 85].

Less obviously but similarly important, Turing chose *not* to inquire into any adaptive function, in Darwinian terms, of the patterns so produced. These patterns may or may not serve an adaptive function, and what that function amounts to is of secondary concern at best. Explaining the generation of their form does not contribute to explaining that form's function, nor does it depend on that function. In this respect, too, Turing's thoughts appear to be in line with, if not explicitly endorsing, D'Arcy Thompson's skeptical view of the relevance of adaptation by natural selection in evolution. The formative processes in organisms are considered at least partly autonomous from Darwinian mechanisms. Whether the florets of a sunflower are patterned on a Fibonacci series, as they in fact are, or whether they are laid out in grid-like fashion, as they possibly *cannot* be according to the mathematical laws of form expounded by Turing, is unlikely to make a difference in terms of selective advantage. In turn however, natural selection may not offer a path to a grid-like pattern in the first place, while enabling, but arguably not determining, the Fibonacci pattern. In likewise fashion, the cognitive abilities of human beings or other animals would not in the first place be considered as adaptive abilities, defined in relation to challenges posed by their environments, but in their, mathematically modellable, form.

Turing's bias towards form over function, in conjunction with his methodological individualism, created a difficulty in systematically grasping a relation that might look straightforward or even obvious to the contemporary reader, who is likely to be familiar with the role of populations and environments in evolution, and who might also be familiar with philosophical concepts of functions: *analogy of functions* across different, phylogenetically distant species. In Turing's notion of decoupling logical form from physical structure, the seeds of the concept of functional analogy appear to be sown, however without growing to a degree of maturity that would prevent the premature conclusions often drawn from Turing's presentation of the TT.

It is the condition of observable similarity in behaviour that has been prone to misguide both proponents and critics of the TT. One cannot straightforwardly deduce a similarity of kind – in this case, being in command of a shared form of intelligence – from a similarity in appearance. A relation of proximity in kind could only be firmly established on the grounds of a relation of common descent, that is, from being part of the same biological population or from being assembled according to a common design or *Bauplan*. This is the ultimate skeptical resource for the AI critic who will never accept some computer's or robot's trait as the same or equivalent to a human one. However convincing it may look to the unprejudiced observer, any similarity will be dismissed as a feat of semi-scientific gimmickry. Even a 1:1 replica of a human being, down to artificial neurones and artificial muscles made of high-tech carbon-based fibres, is unlikely to convince him or her. What the skeptic is asking for is a *structural homology* to lie at the foundation of observable similarities.

In the biological discipline of morphology, the distinction between analogies and homologies has first been systematically applied by Richard Owen, who defined it as follows:

> "ANALOGUE." – A part or organ in one animal which has the same function as another part or organ in a different animal.
> "HOMOLOGUE." – The same organ in different animals under every variety of form and function. [15, p. 7, capitalisation in original]

This distinction was put on an evolutionary footing by Charles Darwin, who gave a paradigmatic example of homology himself, when he asked: "What can be more curious than that the hand of a man, formed for grasping, that of a mole for digging, the leg of the horse, the paddle of the porpoise, and the wing of the bat, should all be constructed on the same pattern, and should include the same bones, in the same relative positions?" [5, p. 434] – where the reference of "the same" for patterns, bones and relative positions is fixed by their common ancestral derivation rather than, for Owen and other Natural Philosophers of his time, by abstract archetypes.

In contrast, an analogy of function of traits or behaviours amounts to a similarity or sameness of purpose which a certain trait or behaviour serves, but which, firstly, may be realised in phenotypically variant form and which, secondly, will not have to be derived from a relation of common descent. For example, consider the function of vision in different species, which is realised in a variety of eye designs made from different tissues, and which is established along a variety of lines of descent. The most basic common purpose of vision for organisms is navigation within their respective environments. This purpose is shared by camera-based vision in robots, who arguably have an aetiology very different from any natural organism. Conversely, the same navigational purpose is served by echolocation in bats, which functions in an entirely different physical medium and under entirely different environmental circumstances, namely the absence of light.

There are no principled limitations as to how a kind of function is realised and by what means it is transmitted. The way in which either variable is fixed depends on the properties of the (biological or technological) population and of the environment in question. In terms of determining its *content*, a function is fixed by the relation between an organism's constitution and the properties of the environment in which it finds itself, and thus by what it has to accomplish in relation to organic and environmental variables in order to prevail. This very relation may be identical despite the constitution of organisms and the properties of the environment being at variance between different species. Perceiving spatial arrangements in order to locomote under different lighting conditions would be a case in point. In terms of the *method* by which a function is fixed, a history of differential reproduction of variant traits that are exposed to the variables of the environment in which some population finds itself will determine the functional structure of those traits. If an organism is endowed with a reproducible trait whose effects keep in balance those environmental variables which are essential to the organism's further existence and reproduction, and if this happens in a population of reproducing organisms with sufficient frequency (which does not even have to be extremely high), the effects of that trait will be their functions.[5]

Along the lines of this argument, an analogy of function is possible between different lines of descent, provided that the environmental challenges for various phylogenetically remote populations are similar. There are no a-priori criteria by which to rule out the possibility that properties of systems with a common descent from engineering processes may be functionally analogous to the traits and behaviours of organisms. In turn, similarity in appearance is at most a secondary consequence of functional analogy. Although such similarity is fairly probable to occur, as in the phenomenon of convergent evolution, it is never a necessary consequence of functional analogy. The similarity that is required to hold between different kinds of systems lies in the tasks for whose fulfilment their respective traits are selected. Structural homology on the other hand does neither require a similarity of tasks nor a similarity of appearance, but a common line of descent from which some trait hails, whatever function it may have acquired later along that line, and whatever observable similarity it may bear

[5] This is the case for aetiological theories of function, as pioneered by [23] and elaborated by [11].

to its predecessor. In terms of providing criteria of similarity that go beyond what can be observed on the phenotypical level, functional analogy trumps structural homology.

4 The Turing Test as Demonstrative vs. Investigative Simulation

On the grounds of the above argument, the apparent under-definition of the epistemical role of the TT owes to an insufficient understanding of the possibilities and limitations of functional analogy in the AI debates: It is either confounded with homological relations, which, as there are no common lines of descent between human beings and computers, results in the TT being rejected out of hand as a test for any possible cognitive ability of the latter. Or analogous functions are considered coextensive with a set of phenotypical traits similar, qua simulation, to those of human beings. Either way, it shows that inferences to possible cognitive functions of the traits in question are not warranted by phenotypical similarity. Unless an analogy of function can be achieved, the charge of gimmickry against the TT cannot be safely defused. If however such an analogy can be achieved, the test itself would not deliver the evidence necessary for properly assessing that analogy, nor would it provide much in the way of a suggestion how that analogy could be traced.

One might be tempted to put the blame for this insufficient understanding of functional analogy on Turing himself – but that might be an act of historical injustice. Firstly, he did not claim functional analogies to be achieved by his simulations. Secondly, some of the linkages between the formal-mathematical models which he developed and more recent concepts of evolution that comprise the role of populations and environments in shaping organic functions were not in reach of his well-circumscribed theory of computation. They were not even firmly in place at the time of his writing. Much of contemporary evolutionary reasoning owes to the Modern Synthesis in evolutionary biology, which was only in the process of becoming the majority view among biologists towards the end of Turing's life.[6]

With the benefit of hindsight however, and with the clarification of matters that it allows, is there any role left for the TT to be played in inquiries into human cognition – which have to concern, first and foremost, the *functions* of human cognition? Could it still function as a simulation of serious scientific value? Or, trying to capture Turing's ultimate, trans-mathematical objective more precisely and restating the opening question of this paper: Could the TT still help to identify the common ground human beings and machines would have to share in order to also share a set of cognitive traits? For modified forms of that test at least, the answer might be positive.

First of all, one should be clear about what kind of simulation the TT is supposed to be. If my reconstruction of Turing's proximate aims is valid, the imitation game was intended as a *demonstrative* simulation of the force and scope of the computational method, with no systematic cognitive intent. By many of its interpreters and critics however, it was repurposed as an *investigative* simulation that, at a minimum, tests for some of the behavioural cues by which people normally discern signals of human intelligence in communication, or that, on a maximal account, test for the cognitive capacities of machines proper.

The notions of demonstrative and investigative simulations are distinguished in an intuitive, prima facie fashion in [16, p. 7 f], but may not always be as clearly discernible as one might hope. Demonstrative simulations mostly serve a didactic purpose, in reproducing some well-known behaviours of their subject matter or "target" in a different medium, so as to allow manipulations of those behaviours' variables that are analogous to operations on the target proper. The purpose of flight simulators for example lies in giving pilots a realistic impression of experience of flying an airplane. Events within the flight simulation call for operations on the simulation's controls that are, in their effects on that simulation, analogous to the effects of the same operations in the flight that is being simulated. The physical or functional structure of an airplane will not have to be reproduced for this purpose, nor, of course, the physical effects of handling or mishandling an in-flight routine. Only an instructive simile thereof is required. I hope to have shown that this situation is similar to what we encounter in the TT, as originally conceived. No functional analogy between simulation and target is required at all, while the choice and systematic role of observable similarities is contingent on the didactic purpose of the simulation.

An investigative simulation, on the other hand, aims at reproducing a selection of the behaviours of the target system in a fashion that allows for, or contributes to, an explanation of that behaviours' effects. In a subset of cases, the explanation of the target's functions is included, too. Here, a faithful mapping of the variables of the simulation's behaviours, and their transformations, upon the variables and transformations on the target's side is of paramount importance. No phenomenal similarity is required, and a mere analogy of effects is not sufficient, as that analogy might be coincidental. Instead, some aspects of the internal, causal or functional, structure of the target system will need to be systematically grasped. To this purpose, an investigative simulation is guided by a theory concerning the target system, while the range of its behaviours is not exhausted by that theory: Novel empirical insights are supposed to grow from such simulations, in a manner partly analogous to experimental practice.[7] I hope to have shown that this is what the TT might seem to aim at, but does not achieve, as there is no underlying theory of the cognitive traits that appear to be simulated by proxy of imitating human conversational behaviour.

An alternative proposal for an investigative role of the TT along the lines suggested above would lie in creating analogues of some of the cognitive functions of communicative behaviour. Doing so would not necessarily require a detailed reproduction of all or even most underlying cognitive traits of human beings. Although such a reproduction would be a legitimate endeavour taken by itself, although probably a daunting one, it would remain confined to the same individualistic bias that marked Turing's own approach. A less individualistic, and perhaps more practicable approach might take supra-individual patterns of communicative interaction and their functions rather than individual minds as its target.

One function of human communication, it may be assumed, lies in the co-ordination of actions directed at shared tasks. If this is so, a modified TT-style simulation would aim at producing, in evolutionary fashion, 'generations' of communicative patterns to be tried and tested in interaction with human counterparts. The general method would be similar to evolutionary robotics,[8] but, firstly, placed on a higher level of behavioural complexity and, secondly, directly incorporating the behaviour of human communicators. In order to allow for some such quasi-evolutionary process to occur, there should not be a reward for the machine passing the TT, nor for the human counterpart revealing the machine's nature. Instead, failures of the machine to effectively communicate with its human counterpart, in re-

[6] For historical accounts of the Modern Synthesis, see, for example, [10, 6].

[7] For this argument on the epistemic role of computer simulations, see [22].

[8] For a paradigmatic description of the research programme of evolutionary robotics, see [14].

lation to a given task, would be punished by non-reproduction, in the next 'generation', of the mechanism responsible for the communicative pattern, replacing it with a slightly (and perhaps randomly) variant form of that mechanism. In this fashion, an adaptive function could be established for the mechanism in question over the course of time. Turing indeed hints at such a possibility when briefly discussing the "child machine" towards the end of [19, pp. 455–460] – a discussion that, in his essay, appears somewhat detached form the imitation game proper.

For such patterns to evolve, the setup of the TT as a game of imitation and deception might have to be left behind – if only because imitation and deception, although certainly part of human communication, are not likely to constitute its foundation. Even on a fairly pessimistic view of human nature, they are parasitic on the adaptive functions of communication, which are more likely to be co-operative.[9] Under this provision, humans and machines would be endowed with the task of trying to solve a cognitive or practical problem in co-ordinated, perhaps collaborative, fashion. In such a situation, the machine intriguingly would neither be conceived of as an instrument of human problem-solving nor as an autonomous agent that acts beyond human control. It would rather be embedded in a shared environment of interaction and communication that poses one and the same set of challenges to human and machine actors, with at least partly similar conditions of success. If that success is best achieved in an arrangement of symmetrical collaboration, the mechanisms of selection of behavioural patterns, the behavioural tasks and the price of failure would be comparable between human beings and machines. By means of this modified and repurposed TT, some of the functions of human communication could be systematically elucidated by means of an investigative simulation. That simulation would establish functional analogies between human and machine behaviour in quasi-evolutionary fashion.

5 Conclusion

It might look like an irony that, where, on the analysis presented in this paper, the common ground that would have to be shared between human beings and machines in order to indicate what cognitive traits they may share, ultimately and in theory at least, is functionally identified, and where the author of that thought experiment contributed to developing the notion of decoupling the function of a system from its physical structure, the very notion of functional analogy did not enter that same author's focus. As indicated in section 4 above, putting the blame on Turing himself would be an act of historical injustice. At the same instance however, my observations about the formalistic and individualistic biases built into Turing's computational method do nothing to belittle the merits of that method as such, as its practical implementations first allowed for devising computational models and simulations of a variety of functional patterns in a different medium, and as its theoretical implications invited systematical investigations into the physical underdetermination of functions in general. In some respects, it might have taken those biases to enter this realm in the first place.

References

[1] Margaret A. Boden, *Mind as Machine: A History of Cognitive Science*, Oxford University Press, Oxford, 2006.

[2] B. Jack Copeland, 'The Turing Test', *Minds and Machines*, **10**, 519–539, (2000).

[3] *The Essential Turing*, ed., B. Jack Copeland, Oxford University Press, Oxford, 2004.

[4] B. Jack Copeland, 'The Church-Turing Thesis', in *The Stanford Encyclopedia of Philosophy*, html, The Metaphysics Research Lab, Stanford, spring 2009 edn., (2009).

[5] Charles Darwin, *On The Origin of Species by Means of Natural Selection. Or the Preservation of Favoured Races in the Struggle for Life*, John Murray, London, 1 edn., 1859.

[6] David J. Depew and Bruce H. Weber, *Darwinism Evolving. Systems Dynamics and the Genealogy of Natural Selection*, MIT Press, Cambridge/London, 1995.

[7] Kurt Gödel, 'Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I', *Monatshefte für Mathematik*, **38**, 173–198, (1931).

[8] David Hilbert and Wilhelm Ackermann, *Grundzüge der theoretischen Logik*, J. Springer, Berlin, 1928.

[9] Andrew Hodges, 'What Did Alan Turing Mean by "Machine"?', in *The Mechanical Mind in History*, eds., Philip Husbands, Owen Holland, and Michael Wheeler, 75–90, MIT Press, Cambridge/London, (2008).

[10] Ernst Mayr, *One Long Argument. Charles Darwin and the Genesis of Modern Evolutionary Thought*, Harvard University Press, Cambridge, 1991.

[11] Ruth Garrett Millikan, *Language, Thought, and Other Biological Categories*, MIT Press, Cambridge/London, 1984.

[12] Ruth Garrett Millikan, *Varieties of Meaning*, MIT Press, Cambridge/London, 2004.

[13] James H. Moor, 'An Analysis of the Turing Test', *Philosophical Studies*, **30**, 249–257, (1976).

[14] Stefano Nolfi and Dario Floreano, *Evolutionary Robotics: The Biology, Intelligence and Technology of Self-Organizing Machines*, MIT Press, Cambridge/London, 2000.

[15] Richard Owen, *On the Archetype and Homologies of the Vertebrate Skeleton*, John van Voorst, Lodon, 1848.

[16] *Philosophical Perspectives in Artificial Intelligence*, ed., Martin Ringle, Humanities Press, Atlantic Highlands, 1979.

[17] D'Arcy Wentworth Thompson, *On Growth and Form*, Cambridge University Press, Cambridge, 2 edn., 1942.

[18] Alan M. Turing, 'On Computable Numbers, with an Application to the Entscheidungsproblem', *Proceedings of the London Mathematical Society*, **s2-42**, 230–265, (1936).

[19] Alan M. Turing, 'Computing Machinery and Intelligence', *Mind*, **59**, 433–460, (1950).

[20] Alan M. Turing, 'The Chemical Basis of Morphogenesis', *Philosophical Transactions of the Royal Society, B*, **237**, 37–72, (1952).

[21] Blay Whitby, 'The Turing Test: AI's Biggest Blind Alley?', in *Machines and Thought*, eds., Peter Millican and Andy Clark, volume 1 of *The Legacy of Alan Turing*, 53–62, Clarendon Press, Oxford, (1996).

[22] Eric B. Winsberg, *Science in the Age of Computer Simulation*, University of Chicago Press, Chicago, 2010.

[23] Larry Wright, 'Functions', *Philosophical Review*, **82**, 139–168, (1973).

[9] For an account of the evolution of co-operative functions, see, for example, [12, ch. 2].