Müller, Vincent C. (2019), 'Measuring progress in robotics: Benchmarking and the 'measure-target confusion'', in Fabio Bonsignorio, et al. (eds.), M*etrics of sensory motor coordination and integration in robots and animals* (Cosmos; Berlin: Springer Nature), 169-79. https://www.springer.com/gp/book/9783030141240 – www.sophia.de - http://www.orcid.org/0000-0002-4144-4957

Measuring progress in robotics: Benchmarking and the 'measure-target confusion'

Vincent C. Müller

Anatolia College/ACT & University of Leeds
www.sophia.de

26th May 2018[*]

Abstract: While it is often said that robotics should aspire to reproducible and measurable results that allow benchmarking, I argue that a focus on benchmarking can be a hindrance for progress in robotics. The reason is what I call the 'measure-target confusion', the confusion between a measure of progress and the target of progress. Progress on a benchmark (the measure) is not identical to scientific or technological progress (the target). In the past, several academic disciplines have been led into pursuing only reproducible and measurable 'scientific' results – robotics should be careful to follow that line because results that can be benchmarked must be specific and context-dependent, but robotics targets whole complex systems for a broad variety of contexts. While it is extremely valuable to improve benchmarks to reduce the distance between measure and target, the general problem to measure progress towards more intelligent machines (the target) will not be solved by benchmarks alone; we need a balanced approach with sophisticated benchmarks, plus real-life testing, plus qualitative judgment.

[*] Acknowledgements: I am grateful to Anne Bajart, Fabio Bonsignorio, Barna Ivantovic , Pedro Lima, Matteo Matteuci, Diego Torricelli and Alan Winfield for comments. My thanks to Nick Bostrom for conversations about intelligence testing and measurement. Discussions about the first version of this paper with the audience at IROS 2015 were very useful.

1. Motivation: Towards benchmarking in robotics

There is progress in robotics, so much is clear, but how much progress is there, in which direction, and how can we evaluate the contribution that a particular piece of research makes to this progress? As in any scientific endeavour, answers to these questions require standards for measuring the state of the art, quantifying progress and contributions to progress – contributions to progress have to follow 'scientific method' of the discipline. Evaluations of such contributions, e.g. in peer review, need ways to evaluate not only 'proper method', but also 'progress beyond the state of the art'.

In robotics, these issues have drawn significant attention in recent years, particularly through the 'Good Experimental Methodology and Benchmarking' Special Interest Group in EURON (since 2006) and the IEEE technical committee on 'Performance Evaluation & Benchmarking of Robotic and Automation Systems' (since 2009, see http://www.ieee-ras.org/performance-evaluation/activities). Between them, they have held not less than 31 workshops at leading robotics conferences since 2006 – for a recent summary see (Bonsignorio & Del Pobil, 2015), for a list, see http://www.heronrobots.com/EuronGEMSig/gem-sig-events. The issues have become more urgent as robotics has become more complex (Antonelli, 2015, p. 1) and continues to move into more complex environments and involve more human-computer interaction (Aly, Griffiths, & Stramandinoli, 2017).

Why are these issues of such importance in this particular field? Measuring scientific progress is a challenge in any discipline, but robotics faces particular difficulties. In a first approximation these are:

1. Robotics is mainly an engineering science, it aims to 'make', so a theory can often only be supported by making and testing, rather than by the classic scientific system of 'theory-prediction-measurement'.
2. The interaction between components (hardware and software), emergent properties, environment and whole system performance is extremely complicated – and not easily isolated.
3. The robot hardware and software used in research is often unique, which makes it difficult to reproduce and compare results, or to identify the contribution the components make to progress on a given task.

The 3rd constraint can be practically quite limiting, as Lier/Wachsmuth/Wrede point out: "Experiment testing, execution and evaluation: Advanced robotics experiments require significant efforts spent on system development, integration testing, execution, evaluation and preservation of results. This is particular costly if many of these tasks are carried out manually. Crucial run time parameters and component configurations are often omitted or not documented properly." (Lier, Wachsmuth, & Wrede, 2014, p. 8) This is, of course, not unique in science: many disciplines cannot conduct experiments easily, e.g. for practical reasons (in geology) or for ethical ones (in medicine). Many engineers cannot do so either: A civil engineer cannot build the railway bridge a few times to test what will happen to it under certain conditions, so they have to resolve to modelling and the testing of components.

I this paper, I will present a clarification of terminology, a diagnosis of the problem as an instance of the 'measure-target confusion', a comparison to other sciences, and then a proposal for a resolution.

2. Initial terminology

2.1. Benchmarking in robotics

The 'benchmark' is originally an expression from land surveying for a mark at a known altitude on a fixed object such as a building or a rock. A standardised angle iron could be fixed at this 'mark' as 'bench' for measuring other altitudes from that point onwards, and measuring 'back' to the benchmark could be used for error detection.

So, 'benchmarking' involves measuring and comparison, as well as usually a quantification of results. To achieve a benchmark, the environment has to be *controlled* (like the fixed object at a known altitude) and the system for measuring has to be *standardised* (like the 'bench' of the surveyor). A condition that can be controlled and standardised (to some degree) then be *replicated* – so benchmarking is a form of the standard *scientific experiment*, where results are achieved in a controlled, standardised condition, and can then be replicated.

In robotics, 'benchmarking' is often used in a wider sense for testing conditions that cannot be very precisely replicated, but where performance can be measured, so we will continue this use. In any case, the control and replication of conditions is a matter of degree.

Performance on a benchmark is performance on the benchmark only – it does not allow induction to performance on a different benchmark or context. So if an industrial robot can set *x* welding points per minute in a controlled environment of a particular factory, then it can do this in the next factory, too, since the conditions can be specified by the manufacturer, and reproduced. But, if a particular autonomous car can drive at a particular speed on a closed racing track under particular conditions, this says nothing about its performance on a different track or under different conditions, e.g. with traffic on the racing track.

To put this in more general terms: Performance on a benchmark is not transferable, it says nothing about a TRL (technological readiness level), it will often not be on a systems level and – crucially as we shall see – it involves no flexibility.

As far as I can tell, there is a lack of technical benchmarks in robotics – and in other fields, these have been extremely useful, e.g. in face recognition or speech-to-text. An experiment in science requires that a particular procedure and the measurement of results is described with sufficient accuracy such that the whole 'experiment' can be replicated at another time or another place. This makes it possible, in principle, for the result to be checked for their accuracy – we don't have to take the word of the researchers. In other words, the ability to replicate (or reproduce) a result is a hallmark of science. What we need to see now is how important benchmarks (experiments) can be in robotics.

2.2. Competitions in robotics

While there is a shortage of benchmarks, there are many competitions in robotics, particularly since the success of RoboCup football – for some links, see (Dias, Althoefer , & Lima, 2016) and https://en.wikipedia.org/wiki/Robot_competition (though this page needs updating as of June 2016). Competitions are typically of whole systems and performed at the same location around the same time – rather like competition events in sports like a world championship. They serve various social functions apart from furthering scientific progress itself, in particular they are useful for public relations.

There are two fundamentally different types of competitions, namely where systems compete a) against other systems or b) on tasks. Furthermore, competitions differ significantly in the degree to which the conditions of the conditions are specified and controlled. Classic RoboCup football is highly specified (with its detailed rulebooks), and thus very narrow – a good performance in the competition says very little about performance under slightly varied conditions (e.g. different lightning or different surface). On the other hand, performance can be compared between different competitions. RoboCup Rescue, on the other hand, deliberately has rather different conditions each time.

The hybrid idea of a 'benchmarking competition' is pursued in the RoCK-In@home and @work competitions (Amigoni et al., 2016): The competitions are sufficiently specified to serve as benchmarks. But if a competition is to be a benchmark it cannot involve competing against another system (that would introduce a non-controlled factor) but only against a task. In that case, the competitions are really benchmarks that are carried out at the same time in the same place. To use an analogy with sports, they are not like a football match but rather like javelin throwing at the athletics championships – which is a competition all right, but the contestants can also compete against each other without meeting at the same place; and they do, for example on who holds the world record. In football, where one competes against an opponent, there is no competition at a distance and there are no 'world records'.

So, there are the two types of competition, against others or against tasks, and the competitions can have more or less controlled conditions. Both types of competitions results in a *partial ordering* of momentary performance – so they are not benchmarks, unless the conditions of a competition against a task are sufficiently controlled to allow reproduction at another time or in another location. Current competitions are not 'real life' scenarios, but controlled to some extent.

Replication is a feature of a technical benchmark but, unlike in an experiment, one would not expect a complex system like a robot to perform identically each time – just as one would not expect a human to run the same distance in the 'Cooper test' at each attempt (distance run in 12 minutes on a tartan track).

3. 'The secure path of a science' through replicable experiment?

3.1. 'Good Experimental Methodology and Benchmarking in Robotics' (GEMSig)

The special interest group (SIG) on "Good Experimental Methodology and Benchmarking in Robotics" in the European Robotics Research Network (EURON) – especially Fabio P. Bonsignorio, Angel P. Del Pobil and John Hallam - has urged for some time now that things need to change: "… the current practice of publishing research results in robotics made it extremely difficult not only to compare results of different approaches, but also to assess the quality of the research presented by the authors. […]" (EURON, 2008)

They urge that things need to change in order to allow for better scientific progress:

> Yet in robotics, artificial intelligence, and automation, the reproduction of result from conference and journal papers, as they are today, is quite often very difficult, if not impossible. This situation is bad for science, as it becomes difficult to objectively evaluate the state of the art in a given field, and also it becomes problematic to build on other people's work, thus undermining one of the basic foundations of scientific progress. (Bonsignorio & Del Pobil, 2015, p. 32; cf. Madhavan, del Pobil, & Messina, 2010)

The proposal is quite clear: Benchmarks and experiments are the way to resolve this problem: "EURON has played an important role by fostering systematic benchmarking and good experimental practice in robotics research." (EURON, 2008)

"The main road to follow the scientific method is to allow the replicability of the experiments." (Antonelli, 2015, p. 3) with his characteristic title "Robotic research: Are we applying the scientific method?". It is characteristic that a recent survey of activities at IROS 2015 "Robot competitions: What did we learn?" (Dias et al., 2016) only mentions positive effects: "The aim is to stimulate innovation more effectively, to meet a defined challenge, and to provide solutions to the problems that matter to roboticists and society" and does not differentiate technical benchmarks from testing or competitions. There are now proposed metrics for many fields, including multi-agent systems (Iantovics, Rotar, & Nechita, 2016).

Scientific method seems a laudable aim, and a task well worth fighting for. However, I wonder whether this is really what we want. There are some examples of academic fields that have tried to take the 'secure path of a science' and ended up making things worse. Perhaps it is useful to look at them. Let me start with my own, though this is clearly far removed from robotics.

3.2. Kant's revolution … for robotics?

Immanuel Kant was planning a revolution for philosophy, a 'Copernican revolution' and this project earned him a position as possibly the most important philosopher of modern times, but also put philosophy on a bad track that stymied its progress for at least a century. (The next attempt at 'scientific philosophy' was waiting, in Vienna Circle positivism and 'analytic philosophy'.)

In order to illustrate this interesting parallel between robotics and philosophy, allow me to quote from his classic *Critique of Pure Reason* (Kant, 1791), in particular the Preface to the 2nd edition (1787), known as 'B':

> Metaphysics … though it is older than all other sciences … has not yet had the good fortune to enter upon the secure path [find the secure step] of a science. (B 15) … and is indeed a merely random groping (B 7)

– this is roughly what the GEMSig say about the current state of robotics. And now Kant compares his discipline to others:

> That logic has already … proceeded upon this sure path is evidenced by the fact that since Aristotle it has not required to retrace a single step. … That logic should have been thus successful is an advantage which it owes entirely to its limitations. (B 8)
>
> … mathematics, among that wonderful people, the Greeks, had already entered upon the sure path of science (B 9)
>
> Natural science was very much longer in entering upon the highway of science. (B 13)

Kant then proposes a method for scientific metaphysics, through *replicable experiment* - and narrowing of scope: "This method, modeled on that of the student of nature, consists in looking for the elements of pure reason *in what admits of confirmation or refutation by experiment.*" (fn. 4).

> … such a gift is not to be valued lightly. For not only will reason be enabled to follow the secure path of a science, instead of, as hitherto, groping at random, without circumspection or self-criticism; our enquiring youth will also be in a position to spend their time more profitably than in the ordinary dogmatism by which they are so early and so greatly encouraged to indulge in easy speculation about things of which they understand nothing, and into which neither they nor anyone else will ever have any insight. (B 19)

… and thus philosophy was saved and has hitherto walked happily the secure path of a science – Not really! The overall experimental method turned out unsuitable. The walk was tried in 'German Idealism' in the 19th Century, failed badly, and then we had a backlash into several directions, with new 'scientific' methods or less scientific ones.

And this is not an isolated incident either: Psychology was captured by the 'scientific' behaviourism (it's slogan was "only observable data!") and had to free itself many decades later. History tried just to say 'what actually happened' (L. v. Ranke), but then found that this is an impossible aim and does not allow it to do its job. Etc. etc.

Perhaps there is a lesson to be learned here? I want to suggest that each of these developments are marked by a confusion between reaching a *target* and reaching a quantifiable *measure* on the path towards that target. This is what I call the 'measure-target confusion'.

4. The measure-target confusion: Benchmarking scientists and people

4.1. Useful and useless benchmarks

Allow me some anecdotal evidence in a first explanation of the phenomenon that I see looming here: In 2011, I discussed with a senior person in robotics funding about the need to measure and demonstrate progress, and I suggested that this is also in the interest of funding agencies. Despite general agreement, the initial comment was "We have benchmarks and demos coming out of our ears", and then they explained that they knew full well that systems that work beautifully in the demonstrations (at project reviews) might not do very much afterwards. – Their suspicion was that 'benchmarks' are just a way to *show* success but that they actually did not signify that success has taken place … clearly a call for high quality benchmarking.

A second piece of anecdotal evidence: At a workshop in 2012, I asked a senior person in speech recognition what they saw as an advantage of their field, in comparison to cognitive systems, and they replied that the existence of benchmarks that everybody knows and everybody tests their systems against has proven an extremely useful tool for their field – but they added that at the same time these benchmarks had stifled progress because people only focus on them, their systems and papers are 'designed to the test' and aspects that may be relevant to the field but are not in the test will be ignored – for example information in video data, such as gestures or facial expression.

4.2. The measure-target confusion

The problem we see here is a common one. We observe a certain social development (here: scientific progress) and then try to see how this can be measured (here: benchmarking). So far, so good. But then we turn the *measure* into a social *target*: We ask for research that improves on benchmarks! This is known to be dangerous because targeting the measure will *change* the social practice itself; the practice that we intended to monitor. This is sometimes called "Campbell's law". In his own formulation: "The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor."(Campbell, 1979, p. 80) A more intuitive formulation is "when a measure becomes a target, it ceases to be a good measure." I will call this the 'measure-target confusion'.

4.3. Example: Progress of individual scientists

One example of the measure-target confusion that will be familiar to readers of academic papers is the benchmarking of success for individual scientists (and then departments and universities). The arguments in favour of such benchmarking are quite the same as the ones we have seen above: It is difficult to tell whether some contribution to science contributes real progress; it is hard to compare the performance of scientists, but we need such comparisons, we need an 'objective evaluation' for scientific success (e.g. for promotion and hiring decisions). So we get quantifiable measures like number of publications, teaching evaluation numbers, funding amount acquired, journal impact factors, overall citation count, h-index, etc. The use of these measures presumably really is preferable over entirely intuitive judgments, but if researchers target the measures ('improve my h-index') instead of the original target (scientific quality and progress) then we have a target with much less value than we started off with and our judgments on quality and progress will become worse. Of course, this problem is well known and leads to efforts to improve on the quantitative methods (the benchmarks), but such improvements on measures will never be a solution to the measure-target confusion. However, even if we recognize that the measure *is* not the target, there is still a very useful discussion to be had on whether a particular measure gets *closer* to the target than another measure.

For research metrics, this problem has been recognised in "The Leiden Manifesto for research metrics" (Hick, Wouters, Waltman, de Rijcke, & Rafois, 2015), where the very first of its ten principles, says: "1. Quantitative evaluation should support qualitative, expert assessment". Quite so … but who will listen to the experts if the performance of a researcher can be boiled down to a single number? Which prospective student will listen to the head of department if the student sees a "ranking" (or UK REF result) that shows them "how good" a university really is?

Another example that will be familiar is formalised project management: Projects are expected to set SMART project milestones, i.e. milestones that are 'specific, measurable, achievable, realistic, time-bound'. (People who have been involved in formal research project evaluation will know this phenomenon.) This *measure* is useful and surely much better than mere intuitive judgment, but, again, the danger looms that reaching the milestone *itself* becomes a *target*, and is 'ticked off' without much care for whether the project is actually moving along well – the measure is confused with the target.

5. Progress towards intelligent systems

For our subject here, a crucial type of measure is that of intelligence of humans to one number, which is reduced to a single number, the IQ. Again, the problem is obvious: Is it a measure of the intended phenomenon, namely human intelligence? Is there a such a property, perhaps 'g' (for 'general intelligence factor') that humans have and that can be measured in one dimension, on a scale that was developed to reflect intellectual development in childhood? This illustrates one central problem: It is crucial how close the measure is to the actual target, how much progress on the measure reflects progress towards the target. With a good measure, progress on measure implies at least some progress towards the target.

Though IQ is supposed to be not something a human can improve, this measure can become a 'target' for artificial intelligence – and indeed, in discussions about progress in artificial intelligence (AI), it is often assumed that progress of AI moves on a one-dimensional axis, and is quantifiable to an extent that one can say a system is twice as intelligent or 'far more' intelligent (Bostrom, 2014; Kurzweil, 2005) – all of this without spending any time on the pesky question what 'intelligence' might be. Some researchers on the progress of AI have avoided this and set a single point of 'measure' namely "Define a 'high–level machine intelligence' (HLMI) as one that can carry out most human professions at least as well as a typical human." (Müller & Bostrom, 2016, p. 556). Note how this is not a benchmark. Another classical measure, the Turing-Test is equally neither a benchmark nor an intuitive measure that is clearly related to the overall target of intelligence (Müller & Ayesh, 2012).

Then there is the tradition of 'cognitive systems', i.e. those to think that artificial systems should, and perhaps must, learn from the intelligent abilities of natural systems – and thus the research on artificial systems can help understanding natural systems. In this tradition, cognitive science and artificial intelligence are still two sides of the same coin, even though they do not expect that cognitive science will find algorithms that can just be implemented on different hardware because the body and environment of the system play central roles (this runs under the label of 'embodiment'). How should one formulate 'benchmarks' for such a system?

Gomila and Müller have summarized the situation, following work in the EUCog network, where they define "We submit that a cognitive system is one that learns from individual experience and uses this knowledge in a flexible manner to achieve its goals." (Gomila & Müller, 2012, p. 456) and thus conclude that "Better systems are those able to deal with increasing degrees of world uncertainty – while allowing for increasing environmental variability (in lighting conditions, distances, sizes, time constraints, …)" (Gomila & Müller, 2012, p. 459). On this basis, they specify 30 measures of progress, none of which are benchmarks – but for all of which benchmarks could be specified. How difficult this can be is quite easy to see if one considers a single relevant dimension, namely 'autonomy' of the agent (cf. Müller, 2012).

Contrast this with benchmarking for the robotics 'multi-annual roadmap; MAR (SPARC, 2015). Here, every "Ability" section has "Ability levels" and every "Technology" section has a component "Benchmarks and Metrics" – only that these sections (5.2.4 ff.) specify no benchmarks, instead they are typically a wish-list with more or less detail on desirable features or performance dimensions, some of which allow for a metrics. In some cases, reference to extant benchmarks in related disciplines is made.

6. Benchmarks are measures, not targets

I conclude that we need to specify an overall target as well as a number of specific targets (both on a systems and on a components level). Then set technical benchmarks and measure progress, but be always aware what the targets were and that *benchmarks are measures, not targets*. In this way, we can avoid false dichotomies and robotics will be, in Kant's words, neither 'merely random groping', nor on 'the secure path of a science'. Various degrees of precision and reproducibility are possible and useful, provided we avoid the 'measure-target confusion'. There is no way to precisely specify progress or to measure it, but there are ways to improve our work. We must let many flowers bloom!

References

Aly, A., Griffiths, S., & Stramandinoli, F. (2017). Metrics and Benchmarks in Human-Robot Interaction: Recent Advances in Cognitive Robotics. *Cognitive Systems Research, 43*, 313-323. doi:http://dx.doi.org/10.1016/j.cogsys.2016.06.002

Amigoni, F., Bastianelli, E., Bonarini, A., Fontana, G., Hochgeschwender, N., Iocchi, L., . . . Schiaffonati, V. (2016). Competitions for benchmarking. *IEEE Robotics and Automation Magazine, 22*(3), 53-61.

Antonelli, G. (2015). Robotic research: Are we applying the scientific method? *Frontiers in Robotics and AI, 2*, 1-4. doi:10.3389/frobt.2015.00013

Bonsignorio, F., & Del Pobil, A. P. (2015). Toward replicable and measurable robotics research. *IEEE Robotics and Automation Magazine, 22*(3), 32-35.

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford: Oxford University Press.

Campbell, D. T. (1979). Assessing the impact of planned social change. *Evaluation and Program Planning, 2*(1), 67-90. doi:http://dx.doi.org/10.1016/0149-7189(79)90048-X

Dias, J., Althoefer , K., & Lima, P. U. (2016). Robot competitions: What did we learn? *IEEE Robotics and Automation Magazine*(1, March), 16-18.

EURON. (2008). Survey and inventory of current efforts in comparative robotics research. *European Robotics Research Network*. Retrieved from http://www.robot.uji.es/EURON/en/index.htm

Gomila, A., & Müller, V. C. (2012). Challenges for artificial cognitive systems. *Journal of Cognitive Science, 13*(4), 453-469. doi:10.17791/jcs.2012.13.4.453

Hick, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafois, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature, 520*, 429-431. doi:10.1038/520429a

Iantovics, L. B., Rotar, C., & Nechita, E. (2016). A Novel Robust Metric for Comparing the Intelligence of Two Cooperative Multiagent Systems *Procedia Computer Science, 96*, 637–644. doi:http://dx.doi.org/10.1016/j.procs.2016.08.245

Kant, I. (1791). *Critique of pure reason* (N. K. Smith, Trans.). London: Palgrave Macmillan 1929.

Kurzweil, R. (2005). *The singularity is near: When humans transcend biology*. London: Viking.

Lier, F., Wachsmuth, S., & Wrede, S. (2014). Modeling Software Systems in Experimental Robotics for Improved Reproducibility: A Case Study with the iCub Humanoid Robot. *Humanoids,* (18-20.11.2014). Retrieved from http://pub.uni-bielefeld.de/luur/download?func=downloadFile&recordOId=2705677&fileOId=2705709

Madhavan, R., del Pobil, A. P., & Messina, E. (2010). Performance Evaluation and Benchmarking of Robotic and Automation Systems.

Müller, V. C. (2012). Autonomous cognitive systems in real-world environments: Less control, more flexibility and better interaction. *Cognitive Computation, 4*(3), 212-215. doi:10.1007/s12559-012-9129-4

Müller, V. C., & Ayesh, A. (Eds.). (2012). *Revisiting Turing and his test: Comprehensiveness, qualia, and the real world* (Vol. 7/2012). Hove: AISB.

Müller, V. C., & Bostrom, N. (2016). Future progress in artificial intelligence: A survey of expert opinion. In V. C. Müller (Ed.), *Fundamental issues of artificial intelligence* (pp. 553-570). Berlin: Springer.

SPARC. (2015). *Robotics 2020: Multi-Annual Roadmap for Robotics in Europe*. Release B 03/12/2015. Retrieved from http://www.eu-robotics.net/