How to Understand 'Superhuman Intelligence' in the Debate about AI?

What can be meant by 'superhuman intelligence' or by 'intelligence explosion' in the debate about AI and its supposedly accelerating development?

1. Meant in one way it might mean *computational acceleration and improved computational complexity measures*. Acceleration and improved computational complexity (for instance the capacity to deal with problems of exponential computational complexity classes), and thus an increased complexity (in length and storage space) of the representations to be successfully processed may allow to solve problems which could not be solved before because of a computational time window of solvability below human capabilities, which cannot catch up with solving problems in too short time spans. Theoretical these problems are still solvable for a human or for a Deterministic Turing Machine (DTM) given the *Church-Turing-Thesis* (CTT), which states that what is intuitively computable is computable by DTMs, and their equivalents. A question to be considered is whether such *quantitative* changes add up in the development of future AI to some *qualitative* change. A proper qualitative change would be one to *supercomputability* (i.e., to computability beyond the 'Turing Limit').

2. It is unclear what might be meant by 'intelligence explosion'. A step to supercomputability seems questionable for at least two reasons:

   i. In a finite, discrete universe infinite advice or precision (as required in some notional models of supercomputable machines like Infinite Advice TMs [cf. Burgin 2005]) are *not available*, so that supercomputability devices can neither be constructed nor do they occur naturally. There are physical limits preventing such supercomputability.

   ii. How could supercomputability ever *evolve* out of ordinary computability, as implemented in our programming languages, their corresponding computer

architecture, and – supposedly – our brains?
(Even a supposed super intelligent agent cannot get to supercomputability using self-improving programming in C or Java.)

3. Whereas reasoning beyond our logic faculty is – by definition – beyond our ken and imagination, it is nonetheless a stumbling block for any theory of 'intelligence explosion' or 'super intelligence' that we by no means can see how the laws of quantificational and propositional logic (the realm of the Turing computable) can ever be superseded. If they are not and there is a logical core to both types of intelligences we are still in the realm of (CTT). In one sense of it the 'proofs'/justifications of (CTT) in principle are, by tying it to a minimal concept of algorithm (cf. Gurevich/Dershowitz 2008), at the same time a *refutation* of any prediction of an arriving super intelligence. Even apart from such 'proofs'/justifications of (CTT) the support for (CTT) seems overwhelming, and even if there are notional machines (like Oracle TMs or Infinite Advice TMs) beyond the 'Turing Limit', the *Physical Church-Turing-Thesis* maintains justifiably that we can never built such machines (cf. Bremer 2008).

4. Any realistic understanding with respect to 'intelligence explosion' thus has to focus on *acceleration and increased representational complexity*. Humans could in this case still simulate the super intelligent reasoning *in principle*, but the qualifier "in principle" now points to the substantial difference between super intelligence and ordinary human intelligence. Humans will substantially lack in performance, thus rendering the 'in principle' available simulations or translations of the super intelligent reasoning futile. As problem solving and creativity depend on working memory, simultaneously weighing options and (probabilistic) anticipation, a super intelligence equipped with advanced working memory and calculation speed might *find* solutions where humans cannot, and may even *invent* solutions and techniques which though understandable – in principle – to humans would have taken too long to be discovered, if ever in terms of human history. In this sense [to be made somewhat more precise in the next two sections] the creation of such a super intelligence is the last invention of mankind.

5. Increases in computing speed face a limit: the speed of light (given current physics is right in this respect). This means that the speed up given with any future AI will be *linear*

with respect to human processing speed. Put otherwise: provided linear slowdown the computations of the new intelligence should be computable by humans. The new intelligence does not compute functions which are uncomputable (e.g., solve some variant of the Halting Problem). The new intelligence will not even be able to solve in manageable time (e.g., before the sun burns out) highly complex problems which are computable (i.e., problems in a computational complexity class like EXP or EXPSPACE), given a large enough instance of that problem. Given the very rough categorizations of computational complexity theory it might very well be that the new intelligence falls into *the same complexity class* like human intelligence.

Nonetheless the difference can be extensive enough for all practical human matters. To put it somewhat vaguely consider the following. [Before that remember that even vague classifications are not arbitrary and cannot be substituted for each other at will.]

The 'average speed of the average city car' is a pretty vague notion. We face a huge variety in cars, and a huge range of their speeds or even average speeds. Nevertheless, in most races (i.e., apart from those consisting exhaustively of narrow bents) a race car will vastly outperform a city car, although the city car can also arrive at the destination of the race.

There is an average speed of human communication – however vaguely this specifies a range of speeds. In case a human can no longer follow the communication speed of the new intelligence a disconnection, a hiatus in intelligence has occurred. One may take this lack of communication speed as one's criterion of 'intelligence explosion' in at least a weak sense.

One can imagine a kind of test, similar to the idea of the Turing Test. A machine or cyborg (or whatever) passes the test for super intelligence if we cannot follow its communication and explanations *in real time*, but understand their content in case of *linear slow down* well enough.

[An analogous argument can be run with respect to working memory space. Even if the new intelligence does not fall in another complexity class than human intelligence the difference may be significant enough for all practical purposes.]

6. Given increased speed a new intelligence, supposedly also equipped with a hugely extended working memory, can survey the solution space and the solution attempts to a problem much quicker (at least by brute force) and, therefore, more comprehensively. Therefore, a new intelligence passing the speed test for being 'super human intelligent' may

solve scientific riddles and puzzles in a time frame substantially sped up in comparison with human scientific progress. If this happens, we have a 'singularity' in the weak sense: We may – in principle – understand and check the results delivered by the new intelligence, but we no longer discover them. In that sense discovery of the new intelligence was our last unaided discovery.
Whether the found solutions can be put into practice is another question depending on further advances in technology.

7. New intelligence in the sense specified need not come about as a conscious program, an android, cyborg, or whatever material or immaterial creature imagined. A human being *augmented* by access to computing devices of the critical strength (i.e., devices which by themselves need not possess consciousness nor all aspects of human intelligence) may with these devices *be* a complex super intelligent system. She can reason beyond the limits of non-augmented human beings. The bottleneck of this weak intelligence explosion and weak singularity rests in the interfaces hooking up the augmented human's brain to the computer equipment. This underlines, again, the importance of research and development in brain-machine interface for the issue.

References

Bremer, Manuel (2008). "A Defense of the Church-Turing-Thesis", in: idem, *Conceptual Atomism and Justificationist Semantics*, Frankfurt a.M. et al., pp. 125-36.

Burgin, Mark (2005). *Super-Recursive Algorithms*. Berlin.

Eden, Amnon H. /Moor, James H./Soraker, Johnny H./Steinhart, Eric (Eds.) (2012). *Singularity Hypotheses*. A Scientific and Philosophical Assessment. Berlin.

Gurevich, Yuri/Dershowitz, Nachum (2008). "A Natural Axiomalization of Computability and Proof of Church's Thesis", *The Bulletin of Symbolic Logic*, 14.

Kurzweil, Ray (2005). *The Singularity Is Near*. New York.

Vinge, Vernor (1993). "The Coming Technological Singularity: How to Survive in the Post-Human Era", *Proceedings: VISION-21 Symposium March 3 0-31, 1993*.

Manuel Bremer