A response to online comments regarding Babylon's research paper, *A comparative study of artificial intelligence and human doctors for the purpose of triage and diagnosis*

Babylon welcomes online comments regarding our recent study, entitled *A comparative study of artificial intelligence and human doctors for the purpose of triage and diagnosis from the wider medical and scientific communities.* Our intention with releasing the results of our pilot study was to promote discussion regarding our preliminary findings from our peers and to offer transparency regarding our methodology.

Whilst we recognise that our findings are preliminary and warrant additional investigation and research, we must dispel several misconceptions evident in some of the online comments received over recent days.

In relation to whether the use of a Bayesian reasoner 'qualifies as AI', this is an interesting point of discussion, but this is perhaps anchored in a rather narrow view on what constitutes AI. We and others (see https://www.nature.com/articles/nature14541, for example) would argue that it is precisely the principled mechanism of updating beliefs in the light of new evidence, whilst also factoring in uncertainty, that defines systems grounded in Bayesian reasoning as providing a theoretical foundation for rational decision making in AI.

A description of the full end-to-end system that powers the app was outside of the scope of the paper. However, we should note that although our paper briefly describes the use of a Bayesian model for reasoning and decision making - to compare triage and differential accuracy against a limited set of human doctors - our AI is not restricted to Bayesian reasoning. Inarguably, the NLP for symptom elucidation (which uses deep neural networks) cannot be disregarded as one of the intelligent elements of the entire AI system. Understanding the patient's input is an essential part of any diagnosis and disambiguating between different contexts itself can be regarded as a test for intelligent agents. A deep neural network is also used (https://arxiv.org/abs/1711.00695) to enable rapid and scalable inference for reasoning on the Bayesian generative model, that is briefly described in the paper.

In addition, the independently created vignettes were *not* used for the model training, development or adjustment in any way -- the cases on which the model was tested were fully unseen to the model making it a perfectly valid performance assessment set. Secondly, none of the doctors used in role-playing were part of the team who tested or developed the AI system, thus shielding them from knowing how the system preferred to ingest information and alienating such biases. Thirdly, as mentioned in the paper, we cover the majority of medical conditions encountered in General Practice in the United Kingdom, which is an extensive list of conditions that have good coverage of what would be observed 'in-the-wild'. Fourthly, with regards to the analysis of the metrics reported: (i) whilst understanding the limitations of both the top 1 and top 3 metrics, we include them for the Semigran vignettes in order to provide a direct comparison. Given that the output of our system is a differential (with associated probabilities) ranked in order of likelihood, using only the top 1 metric would be useless in situations where the most likely diagnosis is not the most important disease to investigate; and (ii) Doctor B and Doctor E were tested on roughly the same number of vignettes, hence it is simply not true that we exploited Doctor B's performance as implied by some commentators. Lastly, we are available to clarify some of the other details related to the experiments conducted to the paper and restate our commitment to uphold the highest scientific rigor to our findings.

Saurabh Johri, Chief Scientist, Babylon
Wednesday 4th July, 2018