This is a pre-print version. Please refer to the final version, published in the *Journal of Consciousness Studies, 30(9-10), 2023*
https://www.ingentaconnect.com/content/imp/jcs/2023/00000030/f0020009

Editorial introduction: Possible introspective systems[1]

François Kammerer (Ruhr-Universität Bochum)
Keith Frankish (University of Sheffield)

Humans *introspect*: they represent their own current mental states in a way that allows for online behavioural control. And the psychological and epistemological specificities of exactly how they do so have long fascinated philosophers and scientists, playing a key role in various metaphysical and methodological debates in the history of thought — many continuing today in contemporary guises.

We believe that, beyond these discussions about the peculiarities of how humans introspect, there is a more general question that is both worth exploring and currently underexplored: *What could introspection be?* What are the various ways in which cognitive systems — human and non-human, natural and artificial, actual and possible — could represent their own mental states in a way that allows for online behavioural control?

This is the question we ask in our paper 'What forms could introspective systems take? A research programme', which forms the target article for this special issue. We give the question a precise formulation, argue for its theoretical importance, and propose an interdisciplinary research programme focused on it. In the process, we provide maps, tools, and directions to help identify and describe the various forms introspective systems could take – to explore the space of possible introspective systems. As with any such proposal, the value of our programme will depend on what further research it enables, encourages, and guides. This in turn will require others to take up and pursue the programme.

As well as our paper, this special issue includes fifteen contributions by philosophers and cognitive scientists, each responding in some way to our proposal. Several contributors (**Carruthers and Masciari**, **Fleming**, **Spener**, **Stoljar**) comment on and criticize our programme, giving us an opportunity to clarify and refine the project. Others (**Dołęga**, **Renero**, **Wu**) discuss particular models or theories of human introspection in the context of our research programme, testing and evaluating the conceptual tools we offer. Importantly, however, most contributors explore some aspect of our titular question. Some look at introspective variation among humans (**Fleming**), including neurodivergent individuals (**Billon**) and Buddhist meditators (**Huebner and Kachru**). Some focus on introspection in non-human animals (**Browning and Veit**, **Englund and Beran**, **Mather and Andrade**), while many discuss introspection in artificial systems (**Browning and Veit**, **Dolega**, **Fleming**, **Long**, **Schwitzgebel and Nelson**). Finally, two contributions take a more speculative perspective and discuss introspection in imaginary minds very different from ours — technologically enhanced humans (**Mandik**) and ancillary artificial minds with an

[1] We thank Valerie Hardcastle (Editor-in-Chief of the *Journal of Consciousness Studies*) and Graham Horswell (Managing Editor) for agreeing to host this symposium in the *Journal of Consciousness Studies*. We also thank all the contributors who agreed to take part, as well as those who reviewed and commented on the contributions.

indiscrete functional organization (**Schwitzgebel and Nelson**). By trying to answer our titular question, these contributors start to implement the research programme we propose, and in doing so, they illustrate its value. We are honoured that so many distinguished researchers took up our challenge.

* * *

To help readers navigate the volume, we shall briefly introduce the contributions that follow our target article. In 'Introspection in the disordered mind and the Superintrospectionitis Thesis', **Alexandre Billon** discusses whether subjects suffering from schizophrenia or depersonalization disorder can be thought of as *better* introspectors than 'normal' subjects, concluding with a negative answer. In 'Studying introspection in animals and AIs', **Heather Browning** and **Walter Veit** outline the most promising approaches to studying introspection in non-human animals and ask whether these approaches can be applied to the study of introspection in AIs. In 'Sub-personal introspection', **Peter Carruthers and Christopher F. Masciari** propose an extension to our research programme, arguing that we should not limit it to globally accessible, 'personal-level' introspective representations, but allow for sub-personal forms of introspection too. In 'Models of introspection vs introspective devices: Testing the research programme for possible forms of introspection', **Krzysztof Dołęga** tests the limits and capacities of our conceptual framework by using it to analyse two models of human introspection — signal detection theory and the metacognitive networks model. In 'Studies of primate metacognition are relevant to determining what form introspection could take in different intelligent systems', **Maisy D. Englund and Michael J. Beran** show that there is good evidence that some nonhuman animals (specifically, primates) can introspect, and they describe how to move forward with future research on animal introspection. In 'Metacognitive psychophysics in humans, animals, and AI: A research agenda for mapping introspective systems', **Stephen M. Fleming** argues for the importance of metacognitive psychophysics in investigating possible introspective systems and summarizes some conclusions of the approach concerning introspection in humans, animals, and AIs. In 'Minds in motion and introspective minds', **Bryce Huebner and Sonam Kachru** explore some conceptions of introspecting and meditating minds within the Buddhist tradition, focusing on early Yogācāra thinkers who saw introspection as a complex reconfigurable process, which can be therapeutically re-shaped by contemplative practices. In 'Introspective capabilities in large language models', **Robert Long** argues that LLMs already possess proto-introspective capacities, suggests how we might train them to develop more advanced forms of introspection, and outlines some ethical issues raised by the prospect of fully introspective LLMs. In 'Sliders', **Pete Mandik** engages in philosophical science-fiction, imagining radical technological enhancements of introspection and self-control and exploring the epistemological, ethical, and social consequences they might have. In 'Can we use the study of introspection to assess decision making and understand consciousness in cephalopods?', **Jennifer Mather and Michaela Andrade** argue that it would be premature to investigate

cephalopod introspection, since we haven't yet settled whether cephalopods possess first-order mental states. They go on, however, to describe a range of cephalopod behaviours that are relevant to the first-order question and might provide a basis for future investigation of introspective capacities. In 'The routes of introspection', **Adriana Renero** seeks to enrich the map of possible introspective systems by distinguishing three 'routes' introspective processes can take in humans — selective, cumulative, and predictive. In 'Introspection in group minds, disunities of consciousness, and indiscrete persons', **Eric Schwitzgebel and Sophie Nelson** describe an artificial ancillary mind, which is indeterminate between a unified mind and a collective of individual minds, and argue that processes occurring in such an entity would be indeterminate between communication and introspection. In 'A framework for self-representational capacities?', **Maja Spener** questions whether our research programme can attain its goals, arguing that mapping the space of possible forms of introspection is unlikely to provide a theoretical framework within which existing accounts of human introspection can be systematically compared. In 'How *not* to identify a research program concerning introspection', **Daniel Stoljar** raises a radical objection to our approach, arguing that introspection in our sense is too liberally defined to constitute a proper object of study and that important definitional restrictions would be needed to make our programme tractable. Finally, in 'On possible and actual human introspection', **Wayne Wu** distinguishes various forms of introspection in humans and uses our map of possible introspective devices to analyse them, thus testing and evaluating the capacity of our framework.

These fifteen contributions are followed by our response, titled 'More possibilities for introspection: Reply to commentators'. We begin by responding to contributors who articulate objections to, or evaluations of, our project. We then turn to contributions that test our framework by applying it to particular views or models of human introspection. We next discuss the numerous contributions that attempt to provide answers to our titular question, and we close by drawing some lessons for our project that have emerged from the symposium.