Foreword

Artificial intelligence (AI), and technological development in general, have been largely off the map of analytic philosophy until recently. Part of the reason for this is no doubt their extrinsic character to the field of philosophy narrowly conceived; broad issues related to social reality have rather been the traditional territory of continental thinking. In the case of artificial intelligence, this situation started to change with the idea and challenge of understanding the human mind by reproducing it. John Searle's (1980) *Minds, Brains, and Programs*, with its famous Chinese room thought experiment about the artificial reproducibility of human intelligence, is one of the most often cited philosophy articles. The question of emulating or even surpassing human mental capacities was taken up by a number of prominent authors in the past decades: David Chalmers, Aaron Sloman, Zenon Phylyshyn, Nick Bostrom, to name but a few.

Another direction from which recent philosophical interest in artificial intelligence has been spurred is that of ethical concerns associated with the surge in the production and use of artificial intelligence. We are finding ourselves in a world where versions of philosophers' wildest fantasies, such as the trolley problem and the experience machine scenario, may come true. Addressing such possibilities, as well as more mundane questions related to the manufacturing, use, and human interaction with different types of AI ahead of time seems to be one of the most important tasks philosophy faces today. The current issue is mostly concerned with such normative questions.

Fabio Tollon's paper asks the questions of whether we should consider machines capable of moral action and moral agency, thus as morally responsible for their actions. Out of the three types of agency (following Johnson and Noorman 2014) he considers, the one which attributes *autonomy* to moral agents seems to be problematic in this regard. Despite the fact that surrogate agency, which may even result in actions with moral consequences, is characteristic of some artificial intelligence systems, these are still guided by human intentions, disqualifying them from any status higher than that of moral *entities*. Autonomy, i.e. the capacity to choose freely how one acts, is strongly tied to the idea that only human beings qualify as moral *agents*. Choosing freely means having "meaningful control" over one's actions. Tollon takes issue with both the engineering and the agential senses of autonomy, claiming that machines should not be called autonomous, as this is not a feature at the level of design, while the moral sense of autonomy comes with too much metaphysical load.

Zsuzsanna Balogh's paper highlights the importance of intersubjectivity in human interaction, drawing on the phenomenology of communication. The author emphasizes the fundamental disanalogy between human-to-human and robot-to-human communication, the latter lacking what she labels "thick intersubjectivity". The users of, e.g. socially assistive robots should be made aware of this fundamental difference, she insists: safeguards should be in place, so that those interacting with such robots can avoid misunderstandings, (intentional or inadvertent) self-deception or misguided emotional attachment.

Tomislav Bracanović addresses the problem of autonomous vehicles' behaviour when lives are at stake. Personal ethics settings (PES) would leave the decision of whether the autonomous car behaves in an egoistic or altruistic manner to the passengers themselves. However, as empirical research suggests, in these circumstances egoistic settings would prevail. Neither deontological nor utilitarian theories would support such settings. The alternative would be government enforced mandatory ethics settings (MES). But is it in the governments' purview to decide who lives and dies on the roads? Again, in Bracanović's view, deontologists and utilitarians alike would object. Is there a third way? Bracanović suggests not having any ethics settings at all for autonomous vehicles would be a more justifiable choice.

As in other areas of life, the automation of government could potentially also lead to huge increases in efficiency and better decisions. But could it be justified? Zsolt Kapelner sets out his stand by arguing that decision-making algorithms operating without human supervision could reasonably be expected to lead to better outcomes for the population, and their use could be even more favourable than democratic rule. Kapelner suggests that traditional objections to this rather radical suggestion, including appeals to public justification, will fail. However, he thinks that rule by algorithm cannot be justified, because it places unacceptable constraints on our freedom.

A general concern about the automatization of scientific discovery is raised by Miklós Hoffmann. Is human involvement a necessary component of scientific achievement, or has this ceased to be the case? Hoffmann casts his vote in the positive and uses Max Weber's stance, who considered specialisation and enthusiasm the essence of scientific discovery. In AI systems, we find both of these components lacking, so – while such systems can assist human scientists in the process of scientific advance in a broad range of ways – they cannot make discoveries on their own.

This volume came together as a result of two research projects and a long-standing collaboration between our home institution, the Institute of Philosophy at the Faculty of Humanities, Eötvös Loránd University (ELTE), and the Department of Sociology and Communication, Budapest University of Technology and Economics (BME), through our co-hosted *Action and Context* workshop series launched in 2018 (putting on 3–7 workshops each semester since). Over the last year, this series included several events on responsibility, deontic logic, ethics that were crucial background to papers in this volume by Balogh and Kapelner. In connection with these events, we are grateful to Tibor Bárány, Gábor Hamp, István Szakadát from BME and László Bernáth, Áron Dombrovszki, Szilvia Finta from ELTE.

Through an ongoing grant, no. K–116191 *Meaning, Communication; Literal, Figurative: Contemporary Issues in Philosophy of Language*, financed by the Hungarian Scientific Research Fund – National Research, Development and Innovation Office (OTKA–NKFIH), launched in 2016, we established the *Budapest Workshop for Language in Action* (LiA, lead by Zvolenszky at ELTE Institute of Philosophy). LiA, originally consisting primarily of philosophers working on language, became instrumental in the recent start of another research group that brought together philosophers of language with philosophers working on moral philosophy, philosophy of mind, ethics and logic: a Higher Education Institutional Excellence Grant (begun in 2018) entitled *Autonomous Vehicles, Automation, Normativity: Logical and Ethical Issues* (at ELTE Institute of Philosophy). We gratefully acknowledge both of these sources of funding.

We wish also to thank the *Hungarian Philosophical Review* for the opportunity to compile an AI-themed issue, and its editor-in-chief's and editors' continued support.

Zsuzsanna Balogh, Judit Szalai, Zsófia Zvolenszky
guest editors from ELTE Institute of Philosophy