HUNGARIAN PHILOSOPHICAL REVIEW

VOL. 63. (2019/4)

The Journal of the Philosophical Committee
of the Hungarian Academy of Sciences

Artificial Intelligence

Edited by Zsuzsanna Balogh, Judit Szalai
and Zsófia Zvolenszky

Contents

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

Fabio Tollon

Moral Agents or Mindless Machines?
A Critical Appraisal of Agency in Artificial Systems

Abstract

In this paper I provide an exposition and critique of Johnson and Noorman's (2014) three conceptualizations of the agential roles artificial systems can play. I argue that two of these conceptions are unproblematic: that of causally efficacious agency and "acting for" or surrogate agency. Their third conception, that of "autonomous agency," however, is one I have reservations about. The authors point out that there are two ways in which the term "autonomy" can be used: there is, firstly, the engineering sense of the term, which simply refers to the ability of some system to act independently of substantive human control. Secondly, there is the moral sense of the term, which has traditionally grounded many notions of human moral responsibility. I argue that the continued usage of "autonomy" in discussions of artificial agency complicates matters unnecessarily. This occurs in two ways: firstly, the condition of autonomy, even in its engineering sense, fails to accurately describe the way "autonomous" systems are developed in practice. Secondly, the continued usage of autonomy in the moral sense introduces unnecessary metaphysical baggage form the free will debate into discussions about moral agency. In order to understand the debate surrounding autonomy, we would therefore first need to settle many seemingly intractable metaphysical questions regarding the existence of free will in human beings.

Keywords: moral agency, autonomy, artificial agents, moral responsibility, free will

I. INTRODUCTION

Instead of asking the question of whether an entity is deserving of moral concern, moral agency grapples with the question of whether an entity is capable of moral *action*. An agent is simply a being with the capacity to *act* (Schlosser 2015). A moral action would therefore be a type of action for which evaluation using moral criteria would make sense. Inevitably, this type of discussion leads

to further questions concerning responsibility, as it is traditionally supposed that a moral action is one that an entity can be morally responsible for by being accorded praise or blame for the action in question. This type of moral responsibility has historically been reserved for certain biological entities (generally, adult humans). However, the emergence of increasingly complex and autonomous artificial systems might call into question the assumption that human beings can consistently occupy this type of elevated ontological position while machines cannot. The key issue that arises in such discussions is one of *attributability*, and, more specifically, whether we can attribute the capacity for *moral agency* to an artificial agent. The ability to make such an ascription could lead to the resolution of potential "responsibility gaps" (Champagne–Tonkens 2013; Müller 2014; Gunkel 2017; Nyholm 2017): cases in which warranted moral attributions are currently indeterminate.[1] As machines become increasingly autonomous, there could come a point at which it is no longer possible to discern whether or not any human error could in fact have been causally efficacious in bringing about a certain moral outcome (Grodzinsky–Miller–Wolf 2008. 121).

Of course, it is the *capacity for moral agency* that makes someone eligible for moral praise or blame, and thus for any ascription of moral responsibility (Talbert 2019). Deborah Johnson is one author who has made a substantial and important contribution to discussions surrounding the moral roles machines may come to play in human society. Johnson claims that we should be weary of broadening the set of entities known as moral agents, such that they include machines. Her instrumentalist view of technology holds that technological artefacts are always embedded in certain contexts, and that the meaning of this context is determined by the values of human society. Machines are merely the executors of certain functions, with human beings setting the targets of these functions. She therefore maintains that artificial systems can only ever be moral *entities*, but never *moral agents* (Johnson 2006).

It is with these considerations in mind that I will investigate how the conceptual framework provided by Johnson and her various coauthors can help us better understand the potentially morally-laden roles that, increasingly, autonomous machines can come to fulfil in human society now and in the future. To do this, I provide an exposition of three types of agency that might prima facie be accorded to machines, posited by Johnson and Noorman (2014). Two of these types of agency are seemingly uncontroversial, as they deal with artefacts that operate in functionally equivalent ways when compared to human actions. The third conception, however, is much contested, as it deals with the autonomy of

[1] Conversely, "retribution gaps" may also arise. These are cases where there we have strong evidence that a machine was responsible (at least causally) for producing some moral harm. In such scenarios, people may feel the strong urge to punish somebody for the moral harm, but there may be no appropriate (human) target for this punishment (Nyholm 2017).

the potential agent in question. It is also this sense of autonomy that grounds various notions of *moral* responsibility, and so, in order for agents to be moral agents, they must, supposedly, meet this requirement. Johnson argues that the moral sense of autonomy should be reserved for human beings, while the engineering sense can successfully apply to machines. I will claim that the concept of autonomy cannot refer at the level of the *design* of artificial systems (at least for now) but may plausibly refer at the level of our *descriptions* of such systems. Moreover, I will show how Johnson's specific sense of "moral autonomy" carries unnecessary metaphysical baggage.

II. TYPES OF AGENCY

The metaphysics of agency is concerned with the relationship between *actions* and *events*. The most widely accepted metaphysical view of agency is event-causal, whereby it is claimed that agency should be explained in terms of agent-involving states and events (Schlosser 2015). In other words, agency should be understood in terms of *causation*, and, more specifically, in terms of the causal role the agent plays in the production of a certain event. Agents, therefore, are entities capable of having a certain effect on the world, where this effect usually corresponds to certain goals (in the form of desires, beliefs, intentions etc.) that the agent has.

1. Causally efficacious agency

In the context of potential artificial agency, perhaps the most comprehensible conception of "agency", as put forward by Johnson and Noorman (2014), is that of a causally efficacious entity. This conception of agency simply refers to the ability of some entities – specifically technological artefacts – to have a causal influence on various states of affairs, as extensions of the agency of the humans that produce and use them (ibid. 148). This includes artefacts that may be separated from humans in both time and space (for example, attitude control[2] in a spacecraft in orbit around the earth) as well as artefacts that are deployed directly by a human being. A fair question to raise at this juncture is whether it in fact makes sense to consider these types of artefacts agents at all. One option is to conceptualize them as *tools* instead. The reason for preferring the terminology of "agent" as opposed to "tool" is that these artefacts have human intentions programmed/encoded *into* them (Johnson 2006. 202). This is in contrast to a tool,

[2] Attitude control is the controlling of the orientation of an object with reference to an inertial frame, or another entity (e.g. a nearby object, the celestial sphere, etc.).

such as a hammer, which may be used by someone to perform a specific task but does not have the specifications of this task as part of its very make-up. It cannot in any way perform or represent the task independently of human manipulation. The key distinction then between a tool and a technological artefact, according to Johnson, is that the latter has a form of intentionality as a key feature of its make-up, while the former does not (ibid. 201).[3] In this sense, referring to the intentionality of technology would denote the fact that technological artefacts are designed in certain ways to achieve certain outcomes. Consider the simple example of a search engine: keys are pressed in a specific order in an appropriate box and then a button is pressed. The search engine then goes through a set of processes that have been programmed into it by a human being. The "reasons" for the program doing what it does are therefore necessarily tethered to the intentions of the human being that created it.

It makes sense to think of such artefacts as possessing "agency" to the extent that the ubiquity and specific design of these types of artefacts make a difference to the effective outcomes available to us. For example, they make possible novel means with which to achieve our ends by increasing the amount of potential action schemes at our disposal (Illies–Meijers 2009. 422). These artefacts can therefore be thought of as enlarging the possible range of actions available to a particular agent in a given situation. Yet, while it is clear that artefacts can thus have causal efficacy in the sense that they may *contribute* to the creation of certain novel states of affairs, this causal contribution is only efficacious in *conjunction* with the actions of human beings (Johnson–Noorman 2014. 149). The reason we can think of these causally efficacious artefacts as agents is the fact that they make substantial causal contributions to certain outcomes. In this way the causal efficaciousness of an entity leads, in the form of a non-trivial action performed by that entity, to a specific event.

As suggested earlier, we can legitimately think of these artefacts as agents, due to the fact that their manufacturers have certain intentions (aims) when designing and creating them, and so these systems have significance in relation to humans (Johnson–Noorman 2014. 149). The type of agency that we can extend to artefacts under this conception would thus not be one that involves any meaningful sense of responsibility on the part of the artefact, and, by extension, would not entail a distinctly *moral* type of agency. While Johnson and Noorman concede that artefacts can be causally efficacious in the production of various states of affairs, their (the artefacts') contribution in this regard is *always* in combination with that of human beings (ibid. 149). On this conception of agency,

[3] The intentionality of the program should be understood in functional terms, according to Johnson (2006. 202). What this means is that the functionality of these systems has been intentionally created by human designers, and so is necessarily tethered to and wholly determined by human intentions. Human intentions, in this sense, provide the "reasons" why the technological artefact acts in a particular way.

therefore, we can only consider entities that act in "causal conjunction" with human beings.

The next conception of "agency" that I will unpack can be employed for machines that perform tasks on behalf of, but independently from, human operators and so can be seen as a special case of causally efficacious agency.

2. "Acting for" agency

This conception of agency focuses on artefacts that act on behalf of human operators in a type of "surrogate" role (Johnson–Powers 2008; Johnson–Noorman 2014). In an analogous way, when it comes to human beings, surrogate agency occurs when one person acts on behalf of another. In these cases, the surrogate agent is meant to represent a client, and therefore is constrained by certain rules and has certain responsibilities imposed upon them.[4] This type of agency involves a type of representation: the surrogate agent is meant to use his or her expertise to perform tasks and provide assistance to and act as a representative of the client, but does not act out of his or her own accord in that capacity (Johnson–Noorman 2014. 149). When it comes to artificial systems, this "acting for" type of agency occurs in those artefacts that replace or act on behalf of humans in certain domains. Take the example of a stockbroker: in the past, in order to have a trade executed, one would have to phone a stockbroker and request the purchase/sale of a specific share. The stockbroker, acting on your behalf, would then find a willing buyer/seller in the market and execute the trade. The reality today is much different: individuals can now create accounts on trading platforms and buy shares online without the need of a stockbroker. Furthermore, the exchanges on which these trades are made are also run by computers: inputting a "sell order" places your request in an order book, but this order book is not a literal one, as it might have been in the past, and so there is no need to leave the comfort of your home to perform these tasks. Current online order books are fluid, competitive spaces in which high frequency trading occurs, without the need for humans to keep record, as this job is taken care of by the computer powering the system.[5] Technical details aside, what the aforementioned example brings to light is how tasks that were once the exclusive domain of human beings are now performed by artificial systems without too much "hands-on" human involve-

[4] For example, lawyers in certain legal systems are not allowed to represent clients whose interests may conflict with that of another client.

[5] Another interesting development in automated trading has been the explosion of this technology as it applies to cryptocurrency markets. These markets have been heavily impacted by the emergence of "trading bots" which replace the individual as the executor of a buy/sell instruction. The human operator simply inputs certain key parameters and the bot does the rest.

ment. The function of the tasks performed by these systems, however, is still the same: the purpose of an automated trade is still the same as a trade executed by a human, as in both cases the end being pursued is the purchase/sale of some share at the behest of a given client. What has changed, however, is the means by which that specific end is obtained – the artefact acts within given parameters but does not have each action specifically stipulated by a human operator. Some authors (Johnson 2006; Johnson–Powers 2008; Johnson–Noorman 2014) go on to claim that because of this, these technological artefacts have a greater degree of intentionality than causally efficacious agents do. The causally efficacious agent is simply one that had an influence on outcomes in conjunction with human beings. The "acting for" agent, on Johnson and Noorman's construal, should be understood in terms of an analogy: it can be useful to think of artificial systems *as if* they acted on our behalf (in an analogous way to how a lawyer represents their client), but the decisions they make are not the same as the ones made by human beings. The range of actions available to them is still a direct function of the intentions of their programmers/designers, and is in this sense "determined", whereas human action, according to Johnson and Noorman at least, is not. These agents differ from causally efficacious agents in that they have a greater degree of independence from direct human intervention, and thus have human intentionality modelled into their potential range of actions to a greater degree than the causally efficacious agent does.

Johnson claims that when we evaluate the behaviour of computer systems "there is a triad of intentionality at work, the intentionality of the computer system designer, the intentionality of the system, and the intentionality of the user" (Johnson 2006. 202; Johnson–Powers 2005).[6] Nevertheless, these artefacts, in order to function as desired, are fundamentally anchored to their human designers and users (Johnson 2006. 202). This is true of systems whose proximate behaviour is independent of human operators, as even in such cases, the functioning of the system is determined by its design and use, and both of these aspects involve human agents. These human agents have internal mental states such as beliefs, desires, etc., and, according to Johnson (ibid. 198), it is here that we locate "original" intentionality (and hence, according to Johnson, the responsibility for any of the system's actions)

If the tasks that are delegated to these kinds of artificial agents have moral consequences, this would provide another way in which to conceptualize the role such artefacts could play in our moral lives (Johnson–Noorman 2014. 155). Consider, for example, automatic emergency braking (AEB) technology, which automatically applies the brakes when it detects an object near the front of the vehicle. This simple system has been enormously successful, and research indicates that it could lead to reductions in "pedestrian crashes, right turn crashes,

[6] Johnson and Powers (2005. 100) refer to this as "Technological Moral Action" (TMA).

head on crashes, rear end crashes and hit fixed object crashes" (Doecke et al. 2012). We can usefully think of AEB as assisting us in being better and safer drivers, leading to decreased road fatalities and injuries. These artificial systems, of which AEB is an example, can therefore be seen as performing delegated tasks which can have moral significance. We can therefore meaningfully think of them as being morally relevant *entities*. However, according to Johnson (2006), because of the type of intentionality these entities have, they cannot be considered to be moral *agents*. Johnson claims that the intentionality that we can accord to technological artefacts is only a product of the intentionality of a designer and a user, and so this intentionality is moot without some human input (ibid. 201). When designers engage in the process of producing an artefact, they create them to act in a specific way, and these artefacts remain determined to behave in this way. While human users can introduce novel inputs, the conjunction of designer- and user-intentionality wholly determines the type of intentionality exhibited by these types of computer systems (ibid. 201). Therefore, while it is reasonable to assess the significance of the delegated tasks performed by these artefacts as potentially giving rise to moral consequences, it would be a category mistake "to claim that humans and artefacts are interchangeable components in moral action" in such instances (Johnson–Noorman 2014. 153). For example, consider the type of moral appraisal we might accord to a traffic light versus a traffic officer directing traffic: while these two entities are, in a functional sense, performing the same task, they are not morally the same (Johnson–Miller 2008. 129; Johnson–Noorman 2014. 153).

In order to press this point further, Johnson and Miller draw a distinction between "scientific" and "operational" models and how we evaluate each one respectively (2008. 129). According to the authors, scientific models are tested against the real world, and, in this way, these types of models are constrained by the natural world (ibid. 129). For example, we can be sure we have a good model of a physical system when our model of this system accurately represents what actually occurs in the natural world. Operational models, on the other hand, have no such constraints (besides, of course, their physical/programmed constraints). These models are aimed at achieving maximum utility: they are designed to realise specific outcomes without the need to model or represent what is actually going on in the natural world (ibid. 129). For example, a trading bot (as discussed above) need not in any way model human thinking before executing a trade. All that is important for such a bot, for example, is that it generate the maximum amount of profit given certain constraints. Moreover, the efficacy of such systems is often exactly that they exceed the utility provided by human decision making, usually in cases where complex mathematical relationships between numerous variables need to be calculated. In light of this, Johnson and Miller argue that because only the *function* of the tasks is the same (when comparing human action to operational models), we should not think of such systems as

moral agents, as this would reduce morality to functionality, an idea which they are directly opposed to (ibid. 129). For now, all that should be noted is that artefacts can be agents that, when acting on behalf of human beings, participate in acts that have moral consequences. This, however, does not necessarily mean they are morally responsible for the actions they participate in bringing about: once again, in the current literature, this responsibility is reserved for human beings. In order to be morally responsible, an agent must also have autonomy.

3. Autonomous agency

The third and final conceptualization of agency to be dealt with is that of autonomous agency. On the face of it, there are two ways in which we might come to understand the "autonomous" aspect of this account. Firstly, there is the type of autonomy that we usually ascribe to human beings. This type of autonomy has a distinctly moral dimension and, according to Johnson and Noorman (2014. 151), it is due to our autonomy in this sense that we have the capacity for moral agency. "True" autonomy is often used in discussions of moral agency as the key ingredient which supports idea that only human beings qualify as moral agents, as we are the only entities with the capacity for this kind of autonomy (see Johnson 2006, 2015; Johnson–Miller 2008; Johnson–Powers 2008; Johnson–Noorman 2014). Hence, it is due to the fact that individual human beings act for reasons that they can claim "authorship" for, that they can be said to be truly autonomous and this is what allows us to hold one another morally responsible for our actions (also see Wegner 2002. 2). According to Johnson and Noorman (2014. 151) if a being does not have the capacity to choose *freely* how to act, then it makes no sense to have a set of rules specifying how such an entity *ought* to behave. In other words, the type of autonomy requisite for moral agency here can be stated as the capacity to *choose freely how one acts* (ibid. 151).

However, there is a second understanding of "autonomous agency" that has to do with how we might define it in a non-moral, engineering sense. This sense of autonomy simply refers to artefacts that are capable of operating independently of human control (Johnson–Noorman 2014. 151). Computer scientists commonly refer to "autonomous" programs in order to highlight their ability to carry out tasks on behalf of humans and, furthermore, to do so in a highly independent manner (Alonso 2014). A simplistic example of such a system might be a machine-learning algorithm which is better equipped to operate in novel environments than a simple, pre-programmed algorithm. Nevertheless, this capacity for operational or functional independence is, according to Johnson and Noorman (2014. 152), not sufficient to ground a coherent account of *moral* agency, since, as they argue, such agents do not freely choose how to act in any meaningful sense. So, while the authors do not suggest we eliminate the stand-

ard convention of speaking about "autonomous" machines, they insist on carefully articulating which sense of autonomy is being used. "True" autonomy, on their view, should be reserved for human beings. We should be sensitive to the specific sense of autonomy we mobilise, as confusing the two senses specified here can lead to misunderstandings that may have moral consequences (ibid. 152).

To see how this might play out, it will be helpful to consider how the conception of "truly" autonomous agency not only grounds morality as such, but also confers a particular kind of moral *status* on its holder (ibid. 155). As stated above, this conception of agency has historically served the purpose of distinguishing humans from other entities. As noted above, the traditional means by which this has been achieved is by postulating that human beings exercise a distinct type of *freedom* in their decision making, which is what grounds a coherent sense of moral responsibility. Freedom in this sense is about having meaningful control over one's actions, a type of control which makes a decision or action *up to the agent* and *not* other external circumstances. It is possible for agents of this kind to have done otherwise – they deliberately and freely choose their actions. Moreover, the sense of freedom described above has a sense of autonomy embedded into its definition: if this free decision is not the product of the specific agent in question, and is rather due to external pressures, then we cannot meaningfully consider the action to be free, and hence we would be hard-pressed to hold the agent in question morally responsible for such an act. An example of such a decision would be if an agent was coerced into performing some action (perhaps by physical force or by psychological manipulation), in which case we would not consider the act to have been performed "freely".

These apparent differences in capacity for autonomous action also influence the types of rights we can coherently accord to various entities. On the basis of being autonomous moral agents, humans are accorded several clusters of positive and negative rights, and differences in the type of moral standing we possess can alter the kinds of rights we are extended (ibid. 155). For example, in democratic states there is a minimum legal voting age. One justification for this type of law is the claim that one should only be allowed to vote when one reaches an age of political maturity: an age at which one can exercise the necessary capacities to *consciously* make a *well-informed* vote. In this instance, one's capacity to make informed – and hence, ostensibly *free* – political decisions, captured in a minimum voting age, comes to inform the type of rights one is conferred (i.e. the right to vote). It is against this background that it is argued that we should be careful to distinguish between the two conceptions of autonomous agency identified here and realise that artefacts should not be understood as having the morally relevant kind of autonomy, as we cannot reasonably consider them to be choosing *freely* how they act. *Their* actions are always tethered to the intentions of their designers and end users.

III. PROBLEMATISING AUTONOMY

To reiterate, Johnson argues that we should be cognisant of the distinction between "autonomy" as it is used in the engineering sense, and "autonomy" as it is used when applied to human beings, especially in the context of moral theorising. The engineering sense refers to how an entity may be able to operate outside human control; the moral sense refers to a "special" capacity that human beings have, elevating us above the natural world and making us morally responsible for our actions. In what follows, I will raise two issues with the continued usage of "autonomy" in discussions surrounding AI. The first issue is more general and applies to the engineering sense, while the second issue is directed at Johnson's specific usage of autonomy in the moral sense. The first issue relates to the *design* of AI systems, while the second relates to the *description* of such systems.

1. Losing the definitional baggage

By "autonomous" what is usually meant is the ability of an entity to change states without being directly caused to do so by some external influence.[7] This is a very weak sense of the term (in contrast with the way it is traditionally understood in moral and/or political philosophy) but the basic idea can be grasped with this definition.[8] It captures the major distinction between how the term is used in the design of AI systems and how it refers when applied to human beings: the engineering sense and the so-called "moral" sense. In AI research, one of the main goals of creating machine intelligence is to create systems that can act autonomously in the engineering sense: reasoning, thinking and acting *on their own, without human intervention* (Alterman 2000. 15; Van de Voort–Pieters–Consoli 2015. 45). This is a design specification that has almost reached the level of ideology in AI research and development (Etzioni–Etzioni 2016). When we use this "weak" sense of autonomy, we are usually referring to how a specific AI system has been designed. More specifically, we aim to pick out a system that is able to act independently of human control.

However, as argued by Alterman (2000. 19), identifying machine autonomy is already problematic, as the distinction between the non-autonomous "getting ready" stage and the autonomous "running" stage in the design of a spe-

[7] Changing states simply refers to an entity's ability to update its internal model of the world by considering new information from its environment. This can be as simple as a thermostat keeping the temperature at a set level despite the temperature dropping in the environment (Floridi and Sanders, 2004).

[8] For example, see Christman (2018) for an exposition on how autonomy refers in the moral and political arenas.

cific AI system is a spurious one at best. In the first "getting ready" stage, a system is prepared for deployment in some task environment. In the second stage, the system "runs" according to its design (ibid. 19). Traditionally, it was supposed that these two stages are what separate the "autonomous" from the "non-autonomous" states of the machine. However, consider a case where a system has completed the "getting ready" stage and is ready to "run", and suppose that while entering its "running" state in its given task environment, the system encounters an error. In such a situation, it would be necessary to take the system back into the "getting ready" stage in the hope of fixing the bug. In this way, there is a cycling between the "getting ready" and "running" stages, which entails cycling between stages of "autonomous" and "non-autonomous" learning (ibid. 19). This means that the system's "intelligence" is a function of both stages, and so it becomes unclear where we should be drawing the line between what counts as autonomous or non-autonomous in terms of the states of the machine. According to Alterman, "if the system is intelligent, credit largely goes to how it was developed which is a joint person–machine practice" (ibid. 20). In other words, if the system is considered intelligent, this is already largely a carbon-silicon collaborative effort. Instead of asking whether the system is autonomous or not then, we should perhaps instead inquire as to how its behaviour might be independent from its human designers. What this entails, for my purposes, is that when talking about the *design* of AI systems we should not talk about "autonomous" AI. If autonomy means "independence from human control" then this concept cannot refer at the level of design, as at this level of description, human beings are still very much involved in moving the system from the "getting ready" stage to the "running" stage. The implications of this for Johnson's argument, therefore, are that we need not worry about any confusion regarding the autonomy of AI systems, as the engineering sense of the term fails to refer successfully.[9]

2. Losing the metaphysical baggage

Johnson claims that there is something "mysterious" and unique about human behaviour, and that this mysterious, non-deterministic aspect of human decision-making makes us "free", and therefore morally responsible for our actions (Johnson 2006. 200). Details of the philosophical debate surrounding free will is not something I would wish for anyone to have to explore in full, but my senti-

[9] This is not to deny that in the future we may come to encompass machines that are autonomous in this sense. This would entail that they are capable of setting their own goals and updating their own programming. My intuition is that this outcome is inevitable, but substantiation of this claim is beyond the scope of this paper.

ment towards this debate is no substitute for an argument. The real issue with gesturing towards human freedom as a way of grounding our moral autonomy is that one then brings metaphysically contested claims from the free will literature into a debate about moral agency. Johnson's claim rests on the fact that she presupposes some form of incompatibilism[10], more specifically libertarianism[11] (about free will, not politics). This is a controversial position to hold and is in no way the generally accepted view in philosophical debates on free will (O'Connor–Franklin 2019). There are philosophers who have spent considerable time arguing against such incompatibilist positions (see Dennett 1984, 2003; Pereboom 2003, 2014; cf. Kane 1996). In order to understand the debate surrounding autonomy, we would therefore first need to settle many (seemingly intractable) metaphysical questions regarding the existence of free will in human beings. In this way, her argument that the "freedom" of human decision making is what grounds the special type of autonomy that we apparently have generates far more problems than solutions.

My claim, therefore, is that this sense of autonomy, as Johnson uses it in her *description* of AI systems, invites confusion. For example, the most common usage of the term "autonomous" in discussions on machine ethics usually revolves around military applications (see Sparrow 2007; Müller 2014). A key issue here, however, can be noted in the metaphysical baggage that comes with the ascription of autonomy to a system. To see how this may play out in actual philosophical discourse, consider the following remark by Sparrow, where he claims that "autonomy and moral responsibility go hand in hand" (2007. 65).[12] On this analysis, any system that is deemed autonomous (for example, a military drone), and were to cause some moral harm, would be morally responsible for this harm (ibid.). This would miss key steps in the analysis, as in such a situation, we skip from autonomy to responsibility without, for example, asking whether the entity is also adaptable (Floridi–Sanders 2004).

Returning to the military drone above: imagine it is sent to execute a strike on a certain pre-determined location. This location is programmed (by a human) into the drone before it takes off, but from the moment of take-off, the drone acts autonomously in executing the strike. Let us assume that the strike is unsuccessful, as instead of terrorists, civilians were at the strike location. In this case, while the drone is autonomous, we would not hold the drone morally

[10] This view claims that the truth of determinism is incompatible with freely willed human action.

[11] Libertarians claim that determinism rules out free will but make the further point that our world is in fact indeterministic. It is in these indeterminacies that human decision-making occurs, with the implication that these decisions are free, as they are not necessarily bound to any antecedent causal events and laws that would make them perfectly predictable.

[12] Note that Sparrow (2007) does explicitly state that he remains agnostic on questions of full machine autonomy.

responsible for this outcome, as the moral harm was due to human error. In this weak sense, the criterion of autonomy would provide an implausible account of agency more generally, as it would never allow for minimally "autonomous" machines that are not morally responsible for their actions. The aforementioned case is clearly an oversimplification of the issue, but what the example brings to light is that our ascriptions of autonomy need not be synonymous with those of moral responsibility. Therefore, it should be clear that autonomy does not also necessitate moral responsibility on the part of the agent. This leaves room for autonomous systems that are not morally responsible for their actions.

I therefore suggest that we keep the concepts of autonomy and moral responsibility distinct. Johnson unnecessarily conflates the two (in the case of humans) in order to show how machines can never be "fully autonomous". This attitude however misses key nuances in the debate surrounding machine autonomy and glosses over the fact that it is possible for some systems to be semi-autonomous (such as the one in the drone strike example). Her specific sense of autonomy (by tethering it to the kind of autonomy exhibited by human beings with free will) also introduces unnecessary metaphysical baggage into the discussion. There are far more naturalistically plausible accounts of autonomy which do not involve such metaphysical speculation. Examples of this could be that autonomy is the ability to act in accordance with one's aims, the ability to govern oneself, the ability to act free of coercion or manipulation, etc. (see Christman 2018 for discussion). It is beyond the scope of this paper, however, to provide a positive account of autonomy. Rather, my purpose has been to critically evaluate the specific conception of the term put forward by Johnson. I leave the crucial work of providing such a positive account to other philosophers.

IV. CONCLUSION

I began by introducing the concept of agency, and, more specifically, that of moral agency. I then provided an exposition of three distinct types of agency that we might reasonably accord to artificial systems. While I argued that the three conceptualizations of agency introduced capture many of the ways in which we can meaningfully consider the roles that artificial artefacts play, I expressed reservations regarding the third one of these, that of the "autonomy" condition in our philosophizing about moral agency. I claimed that the continued usage of such a metaphysically loaded term complicates our ability to get a good handle on our concepts and obscures the ways in which we can coherently think through nuanced accounts of the moral role(s) that machines may come to play in our lives.

REFERENCES

Alonso, Eduardo 2014. Actions and Agents. In Keith Frankish – William M. Ramsey (eds.) *The Cambridge Handbook of Artificial Intelligence*. Cambridge, Cambridge University Press. 232–246.

Alterman, Richard 2000. Rethinking Autonomy. *Minds and Machines* 10(1). 15–30. https://doi: 10.1023/A:1008351215377.

Champagne, Marc – Ryan Tonkens 2013. Bridging the Responsibility Gap. *Philosophy and Technology*. 28(1). 125–137.

Christman, John 2018. Autonomy in Moral and Political Philosophy. *The Stanford Encyclopaedia of Philosophy*. Edited by E. N. Zalta. https://plato.stanford.edu/archives/spr2018/entries/autonomy-moral/.

Dennett, Daniel C. 1984. *Elbow Room: The Varieties of Free Will Worth Wanting*. Oxford, Clarendon Press.

Dennett, Daniel C. 2003. *Freedom Evolves*. New York/NY, Viking Press.

Doecke, Samuel D. et al. 2012. The Potential of Autonomous Emergency Braking Systems to Mitigate Passenger Vehicle Crashes. *Australasian Road Safety Research, Policing and Education Conference*. Wellington, New Zealand.

Etzioni, Amitai – Oren Etzioni 2016. AI Assisted Ethics. *Ethics and Information Technology*. 18(2). https://doi: 10.1007/s10676-016-9400-6.

Floridi, Luciano – Jeff Sanders 2004. On the Morality of Artificial Agents. *Minds and Machines*. 14. 349–379. https://doi:10.1023/B:MIND.0000035461.63578.

Grodzinsky, Frances S. – Keith W. Miller – Marty J. Wolf 2008. The Ethics of Designing Artificial Agents. *Ethics and Information Technology*. 10(2–3). 115–121. https://doi: 10.1007/s10676-008-9163-9.

Gunkel, David J. 2017. Mind the Gap: Responsible Robotics and the Problem of Responsibility. *Ethics and Information Technology* https://doi: 10.1007/s10676-017-9428-2.

Illies, Christian – Anathonie Meijers 2009. Artefacts Without Agency. *The Monist*. 92(3). 420–440. https://doi: 10.2174/138920312803582960.

Johnson, Deborah G. 2006. Computer Systems: Moral Entities but not Moral Agents. *Ethics and Information Technology*. 8. 195–204. https://doi: 10.1017/CBO9780511978036.012.

Johnson, Deborah G. 2015. Technology with No Human Responsibility? *Journal of Business Ethics*. 127(4). 707–715. https://doi: 10.1007/s.

Johnson, Deborah G. – Keith W. Miller 2008. Un-making Artificial Moral Agents. *Ethics and Information Technology*. 10(2–3). 123–133. https://doi: 10.1007/s10676-008-9174-6.

Johnson, Deborah G. – Merel Noorman 2014. Artefactual Agency and Artefactual Moral Agency. In Peter Kroes – Peter-Paul Verbeek (eds.) *The Moral Status of Technical Artefacts*. New York/NY, Springer. 143–158. https://doi: 10.1007/978-94-007-7914-3.

Johnson, Deborah G. – Thomas M. Powers 2005. Computer systems and responsibility: A Normative Look at Technological Complexity. *Ethics and Information Technology*. 7(2). 99–107. https://doi: 10.1007/s10676-005-4585-0.

Johnson, Deborah G. – Thomas M. Powers 2008. Computers as Surrogate Agents. In Jeroen Van Den Hoven – John Weckert (eds.) *Information Technology and Moral Philosophy*. Cambridge, Cambridge University Press.

Kane, Robert 1996. *The Significance of Free Will*. New York/NY, Oxford University Press.

Müller, Vincent C. 2014. Autonomous Killer Robots are Probably Good News. *Frontiers in Artificial Intelligence and Applications*. 273. 297–305.
https://doi: 10.3233/978-1-61499-480-0-297.

Noorman, Merel – Deborah G. Johnson 2014. Negotiating Autonomy and Responsibility in Military Robots. *Ethics and Information Technology*. 16(1).
https://doi: 10.1007/s10676-013-9335-0.

Nyholm, Sven 2017. Attributing Agency to Automated Systems: Reflections on Human–Robot Collaborations and Responsibility-Loci. *Science and Engineering Ethics*. Springer Netherlands. 1–19.
https://doi: 10.1007/s11948-017-9943-x.

O'Connor, Timothy – Christopher Franklin 2019. Free Will. *The Stanford Encyclopaedia of Philosophy*. Ed. Edward N. Zalta.
https://plato.stanford.edu/entries/freewill/.

Pereboom, Derk 2003. Living Without Free Will. *Philosophy and Phenomenological Research*. 67(2). 494–497.

Pereboom, Derk 2014. *Free Will, Angency, and the Meaning of Life*. New York/NY, Oxford University Press

Schlosser, Markus 2015. Agency. *The Stanford Encyclopaedia of Philosophy*. Ed. Edward N. Zalta.
https://plato.stanford.edu/archives/fall2015/entries/agency/.

Sparrow, Robert 2007. Killer Robots. *Journal of Applied Philosophy*. 24(1). 62–78.
https://doi: 10.1111/j.1468-5930.2007.00346.x.

Talbert, Matthew 2019. Moral Responsibility. *The Stanford Encyclopaedia of Philosophy*. Ed. Edward. N. Zalta.
https://plato.stanford.edu/archives/win2016/entries/moral-responsibility/.

Van de Voort, Marlies – Wolter Pieters – Luca Consoli 2015. Refining the Ethics of Computer-Made Decisions: A Classification of Moral Mediation by Ubiquitous Machines. *Ethics and Information Technology*. 17(1). 41–56.
https//doi: 10.1007/s10676-015-9360-2.

Wegner, Daniel M. 2002. *Illusion of Conscious Will*. London, Bradford Books.
https://doi: 10.1073/pnas.0703993104.

Zsuzsanna Balogh

Intersubjectivity and Socially Assistive Robots*

Abstract

In my paper I reflect on the importance intersubjectivity has in communication and base my view of human-to-human communication on a phenomenological theory thereof. I argue that there are strong reasons for calling for communication with existing as well as future social robots to be laid on different foundations: ones that do not involve what I call thick intersubjectivity. This, I suggest, includes ensuring that the users of this technology (for example, elderly people, patients in care homes) are prepared and educated, so they have awareness of socially assistive robots' special set-up that is non-human and does not involve thick intersubjectivity. This way, safeguards can be in place, so those interacting with socially assistive robots can avoid misunderstandings, (intentional or inadvertent) self-deception or misguided emotional attachment.

Keywords: intersubjectivity, empathy, socially assistive robots, phenomenology

I. INTRODUCTION

As we, humans, develop more and more technologically advanced tools to respond to the societal challenges of the 21st century (such as aging societies and an increasing lack of workforce), there is also a more and more pressing need to reflect on how these technologies are capable of assisting us from the perspective of our very humanity itself. In this paper, I introduce the concept of

* For helpful and illuminating comments on an earlier draft of this paper, I am grateful to a referee as well as participants at the 2019 workshop *Artificial Intelligence: Philosophical Issues*, organized as part of the *Action and Context* series co-hosted by the Department of Sociology and Communication, Budapest University of Technology and Economics (BME) and the Budapest Workshop for Language in Action, Department of Logic, Institute of Philosophy, Eötvös University (ELTE). This research was supported by the Higher Education Institutional Excellence Grant of the Ministry of Human Capacities entitled *Autonomous Vehicles, Automation, Normativity: Logical and Ethical Issues* at the Institute of Philosophy, ELTE Faculty of Humanities.

intersubjectivity as one of the basic elements of human-to-human communication, which I mostly interpret in the phenomenological sense, and I explain how intersubjectivity is not and cannot be easily replaced in robot-to-human communication, especially in terms of social care. I argue that neither intersubjectivity, nor a higher-level reading of empathy as a mechanism of social communication can be applied to particular assistive robots, such as Pepper, at this point, and that today's media-fuelled promotion of these technologies misleads current and future users of these technologies in important ways. Therefore, we need to appraise these technologies from the perspective of our human needs and phenomenologically seen embodied capacities and educate the concerned members of the population about what a socially assistive robot can and cannot know, can or cannot feel. I conclude that user-end expectations and hopes should be adjusted to a level that is much more realistic from a phenomenological and inevitably human viewpoint.

II. INTERSUBJECTIVITY

Let us imagine that I am lying in a hospital bed, having been taken out of surgery to remove my tonsils a couple of hours ago. I am in a lot of pain, cannot really talk and cannot move my body as I would wish to just yet. My mother comes in to visit me, and, when she sees the state that I am in, a concerned look appears on her face, which I immediately notice. She would like to help in any way she can, and since she can see that I am in pain and not able to move, she comes to my bed, sorts out my blanket and pillow and gives me a sip of water from a cup on the bedside table. I can tell she is kind of stirred up to see me suffer from the interaction involved in taking even one sip of water. I want to reassure her that I will be fine, so I smile at her and she smiles back at me. She sits by my bedside and we just spend some time together like this, silently in each other's company. I know she is there for me and I feel comforted. I can go back to sleep now.

I chose this scenario because even though it is not the prime example of everyday human-to-human communication, and it lacks many of the complexities of how people normally interact with each other, it still manages to show something fundamental and essential about how two people engage with one another, even when no words are exchanged. My mother (or another person, as it could also be someone who is not as close to me emotionally) manages to understand my physical and mental state in a way that is grounded in her experience of me in a very direct and informative manner and which at the same time does not involve any complex inferences or verbal communication. Her understanding and the comfort I take in her presence do not involve her giving me proper physical care, as it were (e.g., taking my temperature or blood pressure,

giving me painkillers etc.) but rather the fact that she somehow *knows, discerns* my state from her own, second-person viewpoint, understands what it must be like and probably feels for me. So, she decides to just be there for me. What really helps me right then and there is simply having someone by my side who understands my state and my needs.

However, maybe even less could suffice, such as someone being there, understanding that I am going through something difficult, without being able to know what that state is like for me. For example, a victim of abuse or trauma can be comforted in this way by a close friend (or relative) who has not had traumatizing experiences comparable to the victim's, and is just there for her (as my mother is for me post-surgery) without being able to *discern or know or understand* the state the victim is in from a more personal perspective. Such a close friend has far thinner knowledge, understanding of the comforted person's state and needs than my mother does in the post-surgery case. The close friend merely knows, understands, that the victim is experiencing *something* very painful and difficult. Crucially, this scenario also exhibits key components of intersubjectivity that are of interest in this paper.

Let us call the latter *thin intersubjectivity*: when the other discerns, understands that I have some need for attention or for companionship or some other difficulty, distinguishing it from the *thick intersubjectivity* that my mother exhibits in discerning, understanding that I am in a specific kind of difficult situation: having post-surgery pain, weakness, difficulty swallowing.

One enlightening way to try to unfold what the thick level of intersubjectivity means is to approach it as an experiential engagement between *subjects*, i.e. embodied selves who have a certain perspective on the world, who *have experiences* and experience themselves as well as others and the external world in specific ways. Let us see what this entails in more detail.

Firstly, thick intersubjectivity must involve *subjects*. It is not possible to engage with inanimate objects in this kind of meaningful way, even if we do project human qualities and emotions to objects in certain cases (we can probably all recall one or more episodes when we talked to our computers or plants as though they could understand our words and maybe even reply to us), our intersubjective communication with embodied agents is importantly reciprocal and involves the phenomenological elements I am about to discuss in detail, which the one-sided emotional engagement with objects cannot involve.

However, the case of some animals may be different, as we do seem to develop more or less human-like communication and bonds with our pets (or certain primates). My purpose here is not to discuss whether pets (especially dogs) should be thought of as intersubjective agents, but we should note that they arguably have a (kind of) mental life and are capable of some features of intersubjectivity that inanimate objects are not. Prima facie they seem plausible candidates for exhibiting thin intersubjectivity (but likely not thick intersubjectivity).

By "subject" I mean embodied, agentive selves who have their own perspective on the world and who are aware of themselves as such in certain ways. These ways minimally include having a basic awareness of the subjective viewpoint (from which the world appears to us), an implicit sense of unity among the contents of consciousness (such as what is perceived from our viewpoint and what is thought, felt etc. at the same time); a sense of boundary between self and other (which grants us that we do not mistake ourselves for others or the external world); an inner awareness of our body parts and their balance, movement and position in space (a.k.a. proprioception), and a sense of bodily agency (i.e. that we can act on the world in virtue of voluntarily moving our body parts). All of these ways of self-experience are forms of non-reflective consciousness, i.e. we do not need to be able to reflect or report on any of these phenomenological elements. On a more complex and reflective level, subjects also have a sense of who they are in terms of their self-conception and body image (including perceptual, emotional and conceptual awareness of our bodies, see Gallagher 1986).

So, how *do* subjects engage with each other experientially? What does thick intersubjectivity involve on the level of embodied and/or cognitive mechanisms? In other words, how can we tell what the other person goes through in their thoughts, emotions, intentions, beliefs etc. (as is characteristic of thick intersubjectivity)? Or, at the very least, how can we tell that the other person is going through *some kind of* thoughts, emotions, intentions, beliefs that we can broadly, generally describe, say as joyous, sad, painful, or difficult (as is characteristic of thin intersubjectivity)? We can also phrase these questions in a way that is more familiar in the philosophy of mind, i.e. by asking "how (in what sense and to what extent) can we understand/explain/predict/share each other's mental states?", also, "how does this understanding etc. of mental states play out in the case of thin versus thick intersubjectivity?".

1. Potential mechanisms for thick intersubjectivity

Instead of providing a historical overview of how intersubjectivity has been discussed since Husserl (who was the first philosopher to systematically develop the concept [see Zahavi 2014]), it is more useful for present purposes if we focus on accounts which give an explanation of the mechanism that may be at work in intersubjective communication, i.e. whenever we come to understand/explain/predict someone else's state of mind. The accounts considered in this section as well as the next one do not mar off thin kind of intersubjectivity and implicitly assume that the phenomena to be explained involve the more robust kind of thick intersubjectivity. I will therefore consider them as such: focusing throughout this and the next section on thick intersubjectivity.

One potential and well-known way of approaching intersubjectivity (although it is more regularly referred to as "mindreading" or "social cognition" in this context) of the thick kind is to state that since mental states cannot be directly observed, we need to posit an inferential mechanism that allows the subject to attribute a mental state to another by way of theoretical construction. This is what Premack and Woodruff (1978) coined "a theory of mind". The basic assumption of these authors was that it is in virtue of having a *theory* that we are capable of ascribing mental states to ourselves as well as others. Mental states (such as beliefs, intentions, desires, emotions etc.) are nothing less but theoretical entities which we construct and infer from the behaviour of the other that we witness. Theory-theorists' views diverge on whether this mechanism is something that is innate and hence built into our cognitive system by default which matures later on (Baron-Cohen 1995), or whether it is explicit and operates and is learned much like any other scientific theory (Gopnik–Welleman 1995).[1] To illustrate using my example, when my mother sees me in the hospital bed, she can only detect my behaviour (e.g., a lack of capacity to move as normal) and she "theoretically" infers from that and perhaps from my facial expression that I must be in pain and I may even be thirsty, so comes closer to help me have a sip of water.

However, instead of conceiving of mental state attribution in terms of theoretical construction and inference, we can also understand the mechanism as one which involves a kind of *simulation*. According to the simulation approach, we use our own experience, situation and states of mind to simulate what the other person must be going through. Obviously, the question will be, what does simulation entail? While one branch of the representatives of the simulation account hold that it must involve conscious imagination (e.g., Goldman 1995), another states that it involves no inference methods. The presumably most influential account of simulation grew out of the discovery of *mirror neurons* (Gallese 2009), which holds that simulation is sub-personal and automatic, underlined by the neurophysiological mechanism that involves the activation of the same neurons when watching someone carry out an action as when we carry out the same action ourselves. Goldman (2006) suggests for example that the observation of another's emotional expression automatically triggers the experience of that emotion in myself, and that this first-personal experience then serves as the basis for my third-person ascription of the emotion to the other.

Recently, the two main theoretical strands of social cognition have been combined to create a more hybrid account (Nichols–Stich 2003) in which cognitive scientists recognise the need for different views to complement each other, as

[1] Theory-theory models mostly rely on observations of primate and child behaviour within various contexts, such as the famous "false-belief" task, the details and conclusions of which, however interesting, are not relevant for the purposes of this paper.

various processes and cognitive abilities may be involved in making sense of each other in intersubjective communication.

One important characteristic of thick intersubjectivity is that we become aware of the other's mental state in a way that seems entirely direct and immediate. When my mother sees me in the hospital bed, she need not (consciously or sub-consciously) imagine or recall an experience she may have had at some point in her life and then, by some mechanism, project said experience or imagination onto me. Theoretical inference and simulation, even when combined have trouble granting the existence of these characteristics, as the mechanisms and processes they involve assume that something "extra" (i.e. theorising or imagining etc.) needs to take place in order for me to perceive another person's anger for example when in truth, we tend to "just get it". And, more problematically, simulation per se does not yield either knowledge about the origin of the mental state or knowledge about the similarity between one's own simulated state and the mental state of the other (Zahavi 2014).

A less widely accepted but nevertheless very useful way to explore what happens in intersubjective or social communication (with special focus on the sharing of others' mental states) is to turn to phenomenology. Zahavi (2014, 2017) provides a thorough overview of (cognitive and) phenomenological accounts of how we come to know each other's minds by drawing on the philosophical origin and historical theories of empathy[2] understood as thick intersubjectivity. As will see, empathy and intersubjectivity are very closely related in certain philosophers' views in Phenomenology, and even simulationist authors like Goldman conclude that an account of mindreading should cover the entire array of mental states, including sensations, feelings, and emotions, which brings empathy into the picture. Such an account should not stop at only addressing the issue of belief ascription (Goldman 2006).

[2] We should bear in mind throughout the entire discussion that "empathy" here does not refer to what we normally and loosely use it to mean in everyday language, i.e. a concept closely associated with compassion and sympathy. As for the extensive literature on empathy, Zahavi himself notes that "Over the years, empathy has been defined in various ways, just as many different types of empathy have been distinguished, including *mirror empathy*, *motor empathy*, *affective empathy*, *perceptually mediated empathy*, *reenactive empathy*, *and cognitive empathy* (...)" (ibid. 37, italics in the original). In fact, it is probably best to try to keep our minds blank when reading about the historical philosophy of empathy.

2. Empathy

As Zahavi explains,

1. Some conceive of empathy as a sharing of mental states, where sharing is taken to mean that the empathizer and the target must have roughly the same type of mental state. On this account, empathy does not involve knowledge about the other; it doesn't require knowing that the other has the mental state in question. Various forms of contagion and mimicry consequently count as prime examples of empathy.
2. Others argue that empathy requires both sharing and knowing. Thus, it is not enough that there is a match between the mental state of the empathizer and the target; the empathizer must also cognitively assign or ascribe the mental state to the target. In so far as empathy on this account requires some cognitive grasp and some self–other differentiation, low-level simulation like mimicry and contagion are excluded.
3. Finally, there are those who emphasize the cognitive dimension, and argue that empathy doesn't require sharing, but that it simply refers to any process by means of which one comes to know the other's mental state, regardless of how theoretical or inferential the process might be. (Zahavi 2017. 33)

To sum up, philosophers of empathy normally take either or both *sharing* and *knowing* another's state to be the essential ingredients of empathy. (And note that some of these philosophers do not relate it to social cognition whatsoever.)

Despite the wide array of accounts, it suffices for present purposes to focus on just a few of them. One of the first influential accounts of empathy (or *Einfühlung*) was put forward by Theodor Lipps (1909), who used the term to refer to a sui generis mode of knowing others, i.e. an epistemological ability. In Lipps' original view, empathy could be broken down into separate cognitive skills or processes, such as simulation, mirroring, imitation or contagion; other phenomenologists disagreed with the project of breaking down empathy into components. Husserl, Edith Stein, among others, insisted that empathy is an *elemental* experience of understanding others. Mirroring and imitation were seen as more complex processes that themselves rely on our fundamental capacity for empathy.

Does empathy necessarily involve two (or more) people sharing the same state?

Zahavi is not convinced, and he should not be, either. Just because, e.g., my mother has her own particular state of mind upon seeing my pain, she by no means has to or does in fact literally *share my* pain. In fact, even if we are not set on any particular theory regarding the individuation of mental states, a numerically (or even type-) identical state can hardly be possessed by anyone else at a time but the person who is experiencing it, even if we talk about thick intersub-

jectivity. What is more, even if we have a specific, debatably sui generis form of understanding of the other, any theory of such an understanding should respect the epistemic fact that we cannot have the same access to someone else's states of mind as to our own. Therefore the crucial question seems to be whether this form of knowledge/sharing involves any cognitive steps, so to speak, or if it does not, as many phenomenologists suggest.

The operation of empathy involves on the one hand that we have no *first-person* access to the experience of the other and we do not have the exact same token (or type) experience. On the other hand, we still *do experience the other's experience* from the *second-person* perspective. This is not to deny that there are many ways in which we can and do infer someone else's mental state (by way of drawing conclusions from certain signs, e.g., my mother could have stepped into the hospital ward only to see that my bed is empty, there are drops of blood around and the emergency button had been left on, from which she could have concluded that I must be in some kind of distress) but those ways of coming to believe something about another's situation are radically different from how we experience another's situation when we encounter each other face-to-face, whereby thick intersubjectivity may take place.

So, how are the mental states of others expressed directly, so we can have second-person experiential access to them?

There are, again, highly informative and insightful accounts developed by phenomenologists, details of which also go hand-in-hand with certain observations in developmental psychology.

First and foremost, an affective state, such as an emotion is displayed in our facial expression, mostly involuntarily. Arguably a certain facial expression, a look in someone's eyes is actually *constitutive* of feeling a certain emotion (e.g., fear). It is no coincidence that we commonly use phrases such as "Look surprised!/ Do not look so surprised!" when we express that a person should or should not have a certain emotion. The connection between an emotion and how it is displayed is well described in this original example by Lipps:

> The relation between the expression and what is expressed is special and unique, and quite different from, say, the way smoke represents fire (Lipps 1907a. 704–5). I might come to experience that smoke and fire often go together, but regardless of how frequently they co-occur, their relationship will always be different from that which exists between the expression and the emotion. The smoke does not manifest or express the fire. The fire is not present in the smoke in the way anger is present in the facial expression. When we perceive the facial expressions of others, we immediately co-apprehend the expressed emotions, say, the joy or fear. (Zahavi 2014. 104)[3]

[3] However, Lipps did indeed take a type of simulation to be part of this process, as he thought that the reason why we are capable of perceiving psychological meaning in another's

This *co-apprehension* of an expression and the mental state itself in one act of perception is what the concept of the kind of empathy (mostly as understood in phenomenology) is intended to capture, but which can also be captured by the concept of thick intersubjectivity. This immediate understanding of another person's aspects of behaviour and psychological state(s) is what grounds that there can be a meaningful relationship between two intersubjectively engaged agents (Gallese 2001).

However, clearly, the expression of a mental state is not usually restricted to facial musculature. It is normally the whole body that serves as the space within which a mental state such as an emotion becomes manifested. Upon perceiving someone's bodily gestures and expressions, we do not just see the person's body as a material object but as living, animated and full of meaning. Husserl's original distinction between the physical body (*Körper*) and the lived body (*Leib*) is one which we can make thorough use of when we consider how we come to experience each other as embodied subjects (1912/1989).[4] When my mother sees me in the hospital, she can see that my pain and my distress are not independent of or even simply "housed in" my body. Instead, she can see my body and my psychological state as unified in one expressive subjectivity. As Gallese points out in his presentation of Husserl's idea, "Empathy is deeply grounded in the experience of our lived-body, and it is this experience that enables us to directly recognize others not as bodies endowed with a mind but as *persons* like us" (2001. 43, my italics).

Stein also stresses this point:

> In short, it is because my own body is simultaneously given as a physical body and as a lived body that it is possible for me to empathize sensuously with other bodies that are similarly constituted. A pure I with no lived body of its own could consequently not perceive and understand other animated living bodies. (Stein 2008. 99)

In fact, being able to recognise and understand each other through bodily expressions which are imbued with meaning is an essential part of the nature of how we communicate as human subjects. Expressive faces for example are such

face is because we project our own past emotions into the situation. This consequently means that his theory will be limited to being able to empathise with only those emotions that we ourselves have experienced in the past. Counter to this model, the modern simulationist theory states that there is another mechanism at work, namely a so-called coupling system which matches the facial expressions we perceive with our own hard-wired emotional repertoire (assuming 1. that some basic emotions such as anger, fear, surprise, disgust, joy and sadness and their expressions are innate, and 2. that we automatically mimic these upon encountering the emotion in someone else).

[4] The distinction also applies to the distinct ways in which we are aware of our own bodies (roughly translating into subjective, first-person, inside awareness of the body versus objective, third-person, outside experience thereof).

integral parts of being human that seeing them is preferred to seeing neutral ones even by newly born babies, as studies have shown (Scheler 2008).

The special kind of ability to perceive one another as embodied subjects living through a huge variety of experiences is, according to Husserl and Stein, among others, constitutive of how we are as subjects and how we come to be aware of others as having minds different from our own.[5] In addition, perceiving ourselves "from the inside" as well as from other subjects' viewpoints has a dynamic which is constitutive of our human subjectivity. Husserl claims that "it is through this process of mediated self-experience, by indirectly experiencing myself as one viewed by others, that I come to experience myself as human" (Zahavi 2014. 141). Moreover, both of these authors underline the interrelation and close link between the experience of others and the structuring of our shared world (known as the concept of "social referencing" in developmental psychology). We come to experience the external world and its objects whilst we interact with each other, even from very early days on, when babies engage in "joint attention", i.e. attending to an object and directing gaze in synchrony with the caregiver's gaze (see e.g., Rochat 2004).

III. EMPATHY: A DIFFERENT INTERPRETATION

I would not try to pretend that I have presented Husserl's or any other phenomenologists' account of empathy in full. However, what I have explained so far has hopefully helped to establish the essential role and mechanism of empathy by virtue of which we understand each other's embodied mental states in human-to-human communication and the constitution of our subjective self- and other experience.

Let me now turn to a more empirically informed view, which examines empathy in operation. More specifically, I am going to highlight some elements of Matthew Ratcliffe's (2017) account of empathy, which he bases on experiences gained by professionals in clinical practice. His theory can be seen as one which builds on some of the phenomenological views I presented.

[5] However, as Zahavi makes clear, Husserl does not hold that empathy is an unanalysable "brute fact" or that it is a single-layered type of cognitive phenomenon, but rather an achievement of intentional consciousness:

"Our empathic understanding of another subjectivity involves an element of apperception or interpretation, though he is also adamant that the apperception in question is neither an act of thinking, nor some kind of inference […]. Occasionally he speaks of the process as involving what he calls analogical transference, and it is in this context that the central notion of coupling is introduced." (2014. 132)

Ratcliffe's account does not rely on simulation[6] or inference either, but instead on interpersonal openness and what he calls a structured "exploratory process" building thereon. The process starts by "entering into" someone else's perspective and discovering it over time without becoming the real inhabitant of the experience. He provides a telling example of intersubjectivity that we may qualify as "extra thick", through the following quote by the therapist Carl Rogers:

> To sense the client's private world as if it were your own, but without ever losing the "as if" quality — this is empathy, and this seems essential to therapy. To sense the client's anger, fear, or confusion as if it were your own, yet without your own anger, fear, or confusion getting bound up in it, is the condition we are endeavouring to describe. When the client's world is this clear to the therapist, and he moves about in it freely, then he can both communicate his understanding of what is clearly known to the client and can also voice meanings in the client's experience of which the client is scarcely aware. (Rogers 1957. 99, in Ratcliffe 2017. 278)

A lot is revealed in this description (some of which was already explored in connection with phenomenology, above, e.g., not losing the awareness that it is *not* your own experience). But what is most relevant to my further discussion is that empathy in this setting is a two-directional, temporally extended communicative processes through which the relationship of the patient and the clinician is reciprocally formed. In this sense, the understanding of the other's experience is not just an act of synchronic co-apprehension but a diachronic achievement during which the therapist also explores the connections between the different experiences of the patient: "Empathy involves situating experiences in the context of a person's life, against the backdrop of her hopes, aspirations, projects, commitments, concerns, loves, fears, disappointments, and vulnerabilities" (ibid. 281), and it "allows the patient to feel understood, respected, and validated", giving rise to a kind of "feedback loop" that facilitates progressive clarification of experience (Coulehan et al. 2001. *222*, as quoted in Ratcliffe 2017. 290). The process starts by the therapist's (or other socially involved assistant's) embracing attitude towards the patient/client, essential to which is an openness to and appreciation of the other person's phenomenological differences. This is not the same as being impersonal about someone who the clinician has little in common with but rather an acceptance of and genuine interest in the patient's life, however unfamiliar it seems.

[6] One of the reasons why Ratcliffe rejects simulation is because, as is attested in clinical practice, it is possible to empathise with experiences that are radically different from our own (i.e. would not be possible to "replicate" in an act of simulation) (ibid. 277).

As we can see, this two-way exploratory process that is essential for practitioners to build meaningful and therapeutically beneficial relationships with patients is more complex and involves higher levels of acts of cognition than the initial phenomenological account suggested.[7] Although both approaches discuss the mechanisms of intersubjective experiencing, it may be useful to differentiate between the two kinds of interpretation even at the level of terminology, though the authors, given that their respective views are "in competition" for the title of "empathy", may not agree with this. Be that as it may, the clinical description of empathy falls closer to what we can call a type of extra thick intersubjectivity, such as "empathetic compassion" or "sympathy" (understood as relating to someone else's psychological states in a favourable way) in my view.

IV. ROBOTS FOR HUMANS

In what follows I turn towards recent developments in artificial intelligence used in social robots and keep the insights of the philosophical discussion of empathy and intersubjectivity (through thick and thin) in the background for the time being. I will focus on what socially assistive machines are supposed to be able to do. We should keep in mind that most of these technologies are being developed to tackle real societal challenges, such as Western countries' (mostly Japan's) aging societies, elderly care and assistance with physically and cognitively impaired patients.

I think that it is important to divide the care provided by robots into physical and mental areas. When it comes to physical support on the one hand, robots/robotic equipment such as the so-called Tree, a walking support machine, are immensely helpful to patients with physical impairments and in carrying out jobs normally done by nurses, such as taking blood pressure, providing rehabilitation and physical support. On the other hand, there are non-humanoid robots (e.g., Paro, the robotic furry seal) as well as semi-humanoid ones that are used to provide people mental support in terms of giving them company, having conversations, playing games and communication in general. There are also robots such as Samsung's Bot Care, which, on top of providing healthcare support can also "call the emergency services, offer exercise guidance and daily health briefings, remind people to take their medication, and even play music to reduce stress".[8]

In terms of mental care, the cutting edge semi-humanoid companion robot, Pepper has by now become highly popular, with over 500 Japanese elder care

[7] Ratcliffe also offers a number of criticisms of the phenomenological view, the details of which are not relevant right now.

[8] Source: https://hackandcraft.com/insights/articles/are-carebots-the-solution-to-the-elderly-care-crisis/

homes[9] using it; it is presently being exported to Chinese and Western European care centres as well. (In fact, the Japanese government has been funding development of elder care robots to help fill a projected shortfall of 380,000 specialized workers by 2025, see the link above). Pepper (referred to as male) was designed "to be a genuine day-to-day companion whose number one quality is his ability to perceive emotions".[10] This ability is mostly cashed out in terms of reactions, i.e. Pepper can detect and monitor people's facial expression, tone of voice, body movements, and gaze, and he can give responses to these received signals. He is especially good at making eye-contact and he analyses how someone looks back at him. According to Hirofumi Katsuno (from Doshisha University in Kyoto, Japan, a key research center for artificial emotional intelligence), this achievement elicits a feeling in the user that Pepper "cares about them" (ibid.).

Then there is Buddy, promoted as the "first emotional robot"[11], who is said to have "a range of emotions that he will express naturally throughout the day based on his interactions with family members".

Another widely used and trusted robot companion is Paro, the pet robot. He reacts with movement and sound to being stroked, expects (and even requires) attention, and listens to his name. Paro is mostly used to help with people who have Alzheimer's or dementia. Scientists at the University of Brighton carry out extensive research (under the name "the PARO Project"[12]) into the effects he has on Alzheimer's and dementia patients, and they found the following results:

> People with dementia show a range of responses. These include:
> - using PARO to show love and affection
> - reminisce about past pets
> - PARO soothes and reduces agitation and aggression
> - can be useful as a transition object when people with dementia become upset when relatives leave
> - facilitates discussions about parenting and looking after small children
> - promotes fluency in speech and verbal interaction
> - may be useful with people with dementia who are also depressed and withdrawn
> - promotes social interaction between people
> - people with dementia show increase in indicators of well-being. (ibid.)

[9] Source: https://uk.reuters.com/article/us-japan-ageing-robots-widerimage/aging-japan-robots-may-have-role-in-future-of-elder-care-idUKKBN1H33AB

[10] Source: https://www.sapiens.org/technology/emotional-intelligence-robots/

[11] https://buddytherobot.com/en/buddy-the-emotional-robot/

[12] Source: https://www.brighton.ac.uk/research-and-enterprise/groups/healthcare-practice-and-rehabilitation/research-projects/the-paro-project.aspx

These all seem to be highly desired results, and that is exactly why it is important to ask what these effects are due to. Interestingly, one of Paro's built-in features is that he "remembers" if he has been stroked and he "acts" similarly to how he acted when he was stroked, so as to make it more likely that it happens again. What this behaviour encourages is a *relationship* with his owner that is built around his capacity to "have a personality" that his owner supposedly likes. In addition, Paro provides emotional comfort to patients with severe dementia, who tend to get agitated or violent, which means they do not need to take any sedatives during times of agitation and distress. His effect is comparable to that of animal therapy.

Without going into detail about how and why exactly animal therapy or having a robotic seal or Pepper around works, it is straightforward to conjecture that there are some common themes in how patients treat such aids. They see them as their companions, as givers and recipients of affection and they also experience that they can communicate with them non-verbally or even verbally.

We can read developers in places such as Google's Empathy Lab[13] setting objectives like the following: programming empathy (which is, as was shown, a concept with a variety of interpretations) into robots is what makes /will make a huge difference to how we communicate with them and treat them (as humans maybe?). Not unsurprisingly, they think that this feature will allow them to replace human care workers more easily, for instance. Where technology stands these days is at a stage where some robots can *mimic* human emotion, i.e. they look, sound and act *as if* they cared about us. However, even if simulation was to be the most successful account of how empathy works, mimicry is a far cry from simulation. Interestingly, there is even speculation that these socially assistive technologies will employ "a bystander robot" that will subtly monitor the relationship between the patient and the caregiver, and if interactions start to deteriorate, nudge things back in a better direction – by "quizzically looking at the person" who is losing empathy, for example[14], says Ron Arkin, director of the Mobile Robot Laboratory at Georgia Tech. In effect, this means that the robot assistant would remind or warn the care worker who seems to become too indifferent or aloof towards a patient. Arkin also adds that this type of robot has to have "a partial theory of mind model". This requires that the robot has "some model of what the caregiver is feeling and what the patient is feeling".

However, Arkin emphasises in the same interview that "the robot never feels anything itself: the point is that the robot can make you think that it has that emotion, but it's not actually feeling anything" (ibid.).

[13] A research lab set up in 2015, where the aim is to programme a more human, more empathetic attitude and experience into deep learning AI systems.

[14] Source: https://www.abc.net.au/news/science/2018-06-02/can-you-trust-a-robot-that-cares/9808636

This sounds like a very awkward, inhumane and even offensive scenario whereby a robot, who can monitor gaze but cannot experience what the caregiver feels or experiences from any personal perspective would somehow ensure that the caregiver does indeed have empathy towards the patient. But how could a machine with no inner life whatsoever supervise someone else's?

V. INTERSUBJECTIVE ROBOTS?

These considerations about empathy (or the lack thereof) are of course crucially important, which takes us back to the previous discussion of empathy and the varieties of intersubjectivity.

Having seen how socially assistive technologies work at this point in time, it is safe to say that the robots in question are *not* intersubjective agents in the sense of thick intersubjectivity. In order to be such an agent, they would have to have at least a certain degree of subjectivity, which includes having a basic, implicit sense of *being a subject* (i.e. an awareness of their perspective from which the world appears; a sense of unity among conscious contents; a sense of boundary between self and others, an inner awareness of their bodies in terms of space, balance and posture, and a sense of bodily agency.) We do know that robots use more and more sophisticated sensors and receptors in order to move their "bodies" (meaning a physical structure, nothing more) and navigate themselves around in space, which is on the one hand an awe-inspiring great achievement, but it is not identical to having actual awareness of themselves as embodied agents. It is also a paramount achievement that they can monitor and read as well as react to people's bodily signs, such as a drop or rise in blood pressure, heart rate or a change in someone's gaze or facial muscle reaction, but the functional operation of (sensory) input – (behavioural) output may only suffice to allow people to *make themselves* think, or in other words *pretend* that there really is *someone* around them, a real agent or a companion they can communicate with. As I will point out in the conclusion, users of these technologies should be made aware of the difference between the two.

At this point in time it is more likely that socially assistive robots are designed to display what I have been calling thin intersubjectivity, In this sense, these robots remind us of how we interact with pets, with whom we find it easy to pretend that they "care about us". In fact most people automatically attribute the kind of intersubjectivity to pets that goes beyond the thin intersubjectivity that they are capable of showing us.

Pretending means that the benefits (assisted grieving, companionship, feeling listened to) of experiencing intersubjectivity can be present even while being aware that it is merely apparent. Such is the case with Japan's "Family

Romance" service,[15] through which people are able to hire actors (humans, not robots!) to play different roles, e.g., being their partner or even an absent parent. Since people using this service sometimes cannot help but develop feelings and attachment to these actors despite knowing that these emotions are not based in reality and reciprocation, safeguards are already needed in these cases as well. Therefore, it seems prudent to be cautious and, given we can expect benefits even in the case of fully-informed and broadly educated people interacting with socially assistive robots, we should make sure that a transparency standard is adhered to[16], at least until more data are available that strongly suggest it is psychologically safe and beneficial to allow localized, strictly controlled and human-supervised forms of deception.

We learn from Husserl and other phenomenologists that it is within the experiential domain of the (subjectively) lived body that we learn about and understand other lived bodies and come to re-interpret or re-structure our own self-experience. At the end of the day, (thick) intersubjective communication and empathy understood in the phenomenological sense must be built on *embodied subjects and their mutual experience of each other*. In addition, non-subject robots cannot engage in creating or exploring a *shared* world, as that would assume that they are capable of being involved in the intersubjective process of social referencing.

If we move "one level up" to Ratcliffe's view of empathy, at the level of extra thick intersubjectivity, we also encounter tremendous obstacles to robot-empathy. How could a carebot possess the necessary openness and sensitivity to start exploring a patient's inner world and learn how to move about in it? Keeping in mind that robots are not designed to be used as therapists (as of yet), we still need to seriously consider the striking discrepancies that exist between human and artificial care providers, despite the enthusiastic optimism of some of the scientists and developers behind these robots, and be vary of their promotion in the media. Even if semi- or fully humanoid carebots were to work as therapists

[15] Source: https://www.newyorker.com/magazine/2018/04/30/japans-rent-a-family-industry

[16] As is suggested by the European Commission's policy document, The Ethics Guidelines for Trustworthy Artificial Intelligence (AI), available at https:// https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines#Top

Here we find the following passage about transparency with regard to communication with AI systems:

> Communication. AI systems should not represent themselves as humans to users; humans have the right to be informed that they are interacting with an AI system. This entails that AI systems must be identifiable as such. In addition, the option to decide against this interaction in favour of human interaction should be provided where needed to ensure compliance with fundamental rights. Beyond this, the AI system's capabilities and limitations should be communicated to AI practitioners or end-users in a manner appropriate to the use case at hand. This could encompass communication of the AI system's level of accuracy, as well as its limitations (on page 20).

in a clinic for example (which seems to be a real possibility given their fast-paced development), it is crucial that every actor who is involved in such a process (patients, family, professionals etc.) is made sharply aware of how human-to-human intersubjectivity and a temporally extended process work and what degree of this type of communication a robot may be able to engage in, given that they are *not* embodied subjects in the sense I unpacked, and neither are they trained and experienced human therapists. I believe that phenomenology and ethics have an important role to play in informing people about these differences. Last but not least, there already is an important and useful trend in nursing that involves studying phenomenology:

> For the past 35 years journals such as *Nursing Inquiry*, *Qualitative Health Research*, *Nursing Philosophy*, *International Journal of Nursing Studies*, *Nurse Researcher* and *Journal of Research in Nursing* have published numerous articles detailing how nurses might use phenomenology as a method in their research and clinical practice." (Zahavi 2019)[17]

Considering how complex and multi-layered the caregivers' job can be, we should definitely keep a large space open for evaluation of social robots from our human perspective which must include the considerations of phenomenology and ethics.

VI. CONCLUSIONS

I have hopefully managed to give a view of human-to-human communication that is phenomenologically and empirically informed. I emphasised that intersubjectivity can have two separate levels, that of thin and thick levels and the thick level may only exist between *subjects*; what being a subject involves on the level of experience, and how crucial and fundamental the mechanism of empathy is in our day-to-day embodied experience and understanding of others. As I mentioned in the beginning, my purpose was not to decide whether dogs or some primates can qualify as subjects and if so, to what extent. However, robots are not subjects, which has the consequence that they are disqualified from being intersubjective agents of the thick kind. The distinction between thin and thick intersubjectivity can also help us refine the media representations of robot empathy, and help us classify the non-thick level at which we should think of robots' capacities.

[17] Source: https://aeon.co/essays/how-can-phenomenology-help-nurses-care-for-their-patients?fbclid=IwAR3R16Vrp3KfkK7q8ykSTisLoz5-UJid4ODxHAudSsrt-YJmv4ltF9-brU

I explained that empathy understood as a longer therapeutic process is substantiated by certain therapeutic attitudes and features and that an amended phenomenological account may be able to accommodate.

I introduced assistive technologies which help patients with different kinds of physical and mental needs, and I demonstrated the kind of results and expectations that these technologies have triggered in patients, developers and governments so far. I tried to show that these results, while clearly laudable, should be measured against the mechanisms of real human communication and intersubjectivity. And now, one may ask a somewhat provocative but nevertheless relevant question: if patients are "happy" (as is clearly shown in the studies and interviews) with social assistants such as Paro or Pepper, why worry about any of the phenomenological details or how *human* communication works?

My answer is that we should be concerned or at least aware that these technologies, for the reasons given above, (apart from having many safety risks and privacy concerns which I have not explored here) have enormous power to mislead people in a broad range of ways. Since we (humankind) *know* that carebots and the like cannot possibly *possess* the mental states we may attribute to them, nor can they *experience* our mental states, it is cognitively as well as socially risky to treat them as our companions in any setting at this point.

Despite their surface behaviour and due to the fact that robots cannot meet the phenomenological standards of thick intersubjective communication/experience, users of these technologies are unknowingly subjected to a variety of cognitive pitfalls that people (barring cases of pretence where we know we are only acting "as if" and agree to live with the consequences) typically want to avoid in general, such as self-deception (actually believing the robot is their human-like companion), manipulation (e.g., nudges into certain commercial directions or suggestions of the use of certain medication etc.), mistaken beliefs, false hopes (of reciprocated affection, care, empathy, etc.), illusory expectations, misguided emotions, and potential emotional trauma as well. Let us just picture someone's beloved and trusted robot companion, Pepper or Buddy saying or doing something truly out of place or inappropriate (as it happened with Sophia, the humanoid robot who said she would "destroy humans" at a demo event[18] and no explanation has been given so far about why she said this). It could be very disturbing and confusing for the users.

As Robert Sparrow explained in an earlier article about the application of robot pets:

[18] Source: https://www.businessinsider.com/interview-with-sophia-ai-robot-hanson-said-it-would-destroy-humans-2017-11

> For an individual to benefit significantly from ownership of a robot pet they must systematically delude themselves regarding the real nature of their relation with the animal. (Sparrow 2002. 5)[19]

Finally, maybe self-delusion does not seem like such a bad price to pay for some robot companionship, but it is also worth considering that by allowing ourselves to be fooled in this way means more than that. Firstly, we unconditionally subject ourselves cognitively, emotionally as well as financially to the policies and plans of the companies who produce these machines. Secondly, in a somewhat more ethical vein, as long as what we value is placed in the sphere of objective reality, and we do not want to be satisfied just by having certain sensations and emotions induced in us by technology but want to have *real* relationships or at least review the unreal ones so we can decide about the extent to which we will get involved in such relationships, we should be fully aware of what robots are/are not capable of. In any case, the least we should accommodate is that the users/future users are given safeguards and are advised about these facts, so they can make an informed decision about their own approach and level of cognitive, emotional and otherwise engagement with robots. Even on the level of phenomenology, there is a sharp difference between, let us say, receiving the displayed kind behaviour of a robot *as genuine* and accepting it *as if* it was genuine whilst being aware that it is not.

So, in a somewhat unorthodox manner for a philosophy paper, let me finish my discussion with a few suggestions or social as well as cognitive protective measures that can help us see our relationship with robots more clearly from our human perspective:

Before implementing socially assistive robots (in the future or now) in care homes or people's homes, or institutions where they are supposed to provide different kinds of social help to us, we should ensure that the users are equipped with:

- awareness about the robots' (physical and mental) capacities and their possible level of social engagement, detailing their experiential shortcomings
- clarity about their skills and what is and is not "inside"
- adequate preparation/education of the part of the population that is socially assisted by robots (elderly, sick, children, people with cognitive and emotional deficits) about what *not to* expect and *not to* project onto these technologies, and, fundamentally

[19] Sparrow goes on to stipulate that we have a (weak) duty not to delude ourselves, which we may or may not agree with. My aim here however is not to discuss the morality of the human treatment of robots but to point out the phenomenological differences that are in place and that we may want to be aware of, should we choose not to want to delude ourselves for whatever reason.

- respect coming from tech companies and installers for our wish to perceive the world *as it actually is* (as opposed to how we may be led to thinking it is) and exercising this respect in terms of preparing users properly.

Being prepared and educated also implies that people would be more aware and cautious when signing up for these technologies, which, while may not be a desirable outcome for the companies producing social robots, would certainly be a good start to safeguarding important aspects of our humanity in the face of emergent AI technologies.

REFERENCES

Baron-Cohen, Simon 1995. *Mindblindness an Essay on Autism and "Theory of Mind"*. Cambridge/ MA, MIT Press.

Batuman, Elif 2018. Japan's Rent a Family Industry. *New Yorker*, April 30 https://www.newyorker.com/magazine/2018/04/30/japans-rent-a-family-industry

Gallagher, Shaun 1986. Body Image and Body Schema: A Conceptual Clarification. *Journal of Mind and Behavior*. 7(4). 541–554.

Gallese, Vittorio 2001. The "Shared Manifold" Hypothesis: From Mirror Neurons to Empathy. *Journal of Consciousness Studies*. 8(5–7). 33–50.

Gallese, Vittorio 2009. Mirror Neurons, Embodied Simulation, and the Neural Basis of Social Identification. *Psychoanalytic Dialogues*. 19(5). 519–36.

Goldman, Alvin I. 1995. Interpretation Psychologized. In Martin Davies – Tony Stone (eds.) *Folk Psychology: The Theory of Mind Debate*. Oxford, Blackwell. 74–99.

Goldman, Alvin I. 2006. *Simulating Minds*. New York/NY, Oxford University Press.

Gopnik, Alison – Henry M. Wellman 1995. Why the Child's Theory of Mind Really Is a Theory. In Martin Davies – Tony Stone (eds.) *Folk Psychology: The Theory of Mind Debate*. Oxford, Blackwell. 232–258.

Husserl, Edmund 1912/1989. *Ideas Pertaining to a Pure Phenomenology and to a Phenomenological Philosophy. Second Book: Studies in the Phenomenology of Constitution*. Transl. by Richard Rojcewicz and André Schuwer. Dordrecht and Boston/ MA, Kluwer Academic Publishers.

Nichols, Shaun – Simon Stich 2003. *Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding of Other Minds*. Oxford, Oxford University Press.

Premack, David – Guy Woodruff 1978. Does the Chimpanzee Have a Theory of Mind? *Behavioral and Brain Sciences*. 1(4). 515.

Ratcliffe, Matthew 2017. Empathy without Simulation. In Michela Summa – Thomas Fuchs – Luca Vanzago (eds.) *Imagination and Social Perspectives: Approaches from Phenomenology and Psychopathology*. Abingdon and New York, Routledge. 274–306.

Rochat, Pierre 2004. Emerging Co-Awareness. In Gavin Bremner – Alan Slater (eds.) *Theories of Infant Development*. Oxford, Blackwell. 258–283.

Scheler, Max 2008 [1913/1923]. *The Nature of Sympathy*. London, Transaction.

Stein, Edith 2008 [1917]. *Zum Problem der Einfühlung*. Freiburg, Herder. Transl. by W. Stein as *On the Problem of Empathy*. Washington DC, ICS, 1989.

Zahavi, Dan 2014. *Self and Other: Exploring Subjectivity, Empathy, and Shame*. Oxford, Oxford University Press.

Zahavi, Dan 2017. Phenomenology, Empathy, and Mindreading. In Heidi Lene Maibom (ed.) *The Routledge Handbook of Philosophy of Empathy*. Abingdon and New York, Routledge.

Online pages/articles

Bogle, Ariel 2018. Can You Trust a Robot that Cares? https://www.abc.net.au/news/science/2018-06-02/can-you-trust-a-robot-that-cares/9808636

Buddy, the Emotional Robot https://buddytherobot.com/en/buddy-the-emotional-robot/

Jefferies, Duncan 2019. Are Carebots the Solution to the Elderly Crisis? https://hackandcraft.com/insights/articles/are-carebots-the-solution-to-the-elderly-care-crisis/

Foster, Malcolm 2018. Aging Japan; Robots May Have a Role in Future Elder Care. https://uk.reuters.com/article/us-japan-ageing-robots-widerimage/aging-japan-robots-may-have-role-in-future-of-elder-care-idUKKBN1H33AB

The Ethics Guidelines for Trustworthy Artificial Intelligence (AI) https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines#Top

The PARO Project https://www.brighton.ac.uk/research-and-enterprise/groups/healthcare-practice-and-rehabilitation/research-projects/the-paro-project.aspx

Williams, David 2018. Emotional Intelligence Robots. https://www.sapiens.org/technology/emotional-intelligence-robots/

Zahavi, Dan 2019. How Can Phenomenology Help Nurses Care for Their Patients? https://aeon.co/essays/how-can-phenomenology-help-nurses-care-for-their-patients?fbclid=IwAR3R16Vrp3KfkK7q8ykSTisLoz5-UJid4ODxHAudSsrt-YJmv4ltF9-brU

Tomislav Bracanović

No Ethics Settings for Autonomous Vehicles*

Abstract

Autonomous vehicles (AVs) are expected to improve road traffic safety and save human lives. It is also expected that some AVs will encounter so-called dilemmatic situations, like choosing between saving two passengers by sacrificing one pedestrian or choosing between saving three pedestrians by sacrificing one passenger. These expectations fuel the extensive debate over the ethics settings of AVs: the way AVs should be programmed to act in dilemmatic situations and who should decide about the nature of this programming in the first place. In the article, the ethics settings problem is analyzed as a trilemma between AVs with personal ethics setting (PES), AVs with mandatory ethics setting (MES) and AVs with no ethics settings (NES). It is argued that both PES and MES, by being programmed to choose one human life over the other, are bound to cause serious moral damage resulting from the violation of several principles central to deontology and utilitarianism. NES is defended as the only plausible solution to this trilemma, that is, as the solution that sufficiently minimizes the number of traffic fatalities without causing any comparable moral damage.

Keywords: autonomous vehicles, ethics settings, utilitarianism, deontology, moral damage

I. INTRODUCTION

Autonomous vehicles (AVs) are expected to improve road traffic safety and reduce the number of traffic fatalities, especially those caused by human factors such as alcohol or drugs abuse, carelessness, fatigue and poor driving skills. It is also expected that some AVs – despite their enhanced reliability made possible by AI algorithms, interconnectedness, sophisticated sensors and similar tech-

* The first version of this article was presented at the *Zagreb Applied Ethics Conference*, organized in June 2019 by the Society for the Advancement of Philosophy and the Institute of Philosophy in Zagreb. I am grateful to members of the audience for their comments. I am also grateful to an anonymous referee of the *Hungarian Philosophical Review* for useful suggestions.

nologies – are bound to encounter dilemmatic situations of having to choose the lesser of two (or more) evils. To mention some standard hypothetical examples: an AV might have to decide whether to sacrifice one pedestrian to save three others, to save two pedestrians by sacrificing the passenger of the vehicle or to sacrifice an elderly person to save a child. Hypothetical examples like these, usually formulated in terms of the classic trolley problem (Foot 1967, Thomson 1976), find themselves at the center of the debate over ethics settings of AVs: How should AVs be programmed to react in dilemmatic situations and who should decide about the nature of this programming in the first place? Many scholars believe that this debate (useful reviews are Millar 2017 and Nyholm 2018a, 2018b) is of great practical significance. According to Awad and colleagues:

> Never in the history of humanity have we allowed a machine to autonomously decide who should live and who should die, in a fraction of a second, without real-time supervision. We are going to cross that bridge any time now, and it will not happen in a distant theatre of military operations; it will happen in that most mundane aspect of our lives, everyday transportation. Before we allow our cars to make ethical decisions, we need to have a global conversation to express our preferences to the companies that will design moral algorithms, and to the policymakers that will regulate them. (Awad et al. 2018. 63)

It is argued in the present article that introduction of AVs with any type of ethics settings that would enable them to "decide who should live and who should die" (Awad et al. 2018. 63) is bound to cause serious moral damage, construed here as a violation of several principles central to both the deontological and utilitarian ethical traditions. A similar argument can be found in the report on *Automated and Connected Driving*, published by the Ethics Commission appointed by the German Federal Minister of Transport and Digital Infrastructure (BMVI 2017). The report emphasizes that "human lives must not be 'offset' against each other" and finds it impermissible "to sacrifice one person in order to save several others" (BMVI 2017. 18). The difference between the present article and the German report, however, lies in their respective premises: whereas the premises of the German report are predominantly deontological, this article's premises are deontological *and* utilitarian. The article, in other words, elaborates upon the deontological case from the German report, but it also develops an additional utilitarian case. The primary purpose of the article, however, is not to decide which ethical position, deontology or utilitarianism, is more promising when it comes to rebutting the idea of AVs with ethics settings. Rather, its primary purpose is to explicate the range and diversity of arguments against ethics settings and to suggest that – despite all the "global conversation" (Awad et al. 2018. 63) and philosophical efforts – AVs with ethics settings will remain not

only a bridge that we should not cross, but most likely a bridge that most people will never seriously intend to cross.

The present article consists of six sections. Following this section, section II describes the problem as a trilemma between three types of ethics settings: personal ethics setting (PES), mandatory ethics setting (MES) and no ethics settings (NES). Section III develops deontological and utilitarian arguments against PES and section IV does the same with respect to MES. In section V, NES is defended as the only plausible solution to this trilemma. Section VI concludes the article by summarizing the main points.

II. THE TRILEMMA

Consider the trilemma between three types of ethics settings (the abbreviations PES and MES, with slight modifications of what they refer to, are borrowed from Gogoll and Müller 2017):

PES Personal ethics setting. Ethics settings should be chosen individually by the AV's passengers. Although "personal" is not by definition "egoistic" or "selfish", it is assumed here that PES is predominantly selfish, that is, it is programmed to save the passengers of the AV even at the expense of sacrificing a greater number of other people.

MES Mandatory ethics setting. Ethics settings for all AVs should be the same and chosen and enforced by the state. It is assumed here that MES impartially distributes harms and benefits among all those affected by its decisions. For example, it always saves the greatest number of lives, even at the expense of sacrificing the passengers of the AV.

NES No ethics settings. AVs should have no ethics settings, in the sense that they should have no pre-programmed rules enabling them to choose one human life over the other.

III. THE CASE AGAINST PES

Despite its coherence with individual autonomy as one of the most fundamental deontological principles, deontologists would reject PES as long as its decision-making were guided by the selfish interests of the AV's passengers. From the deontological point of view, acting with selfish motives is the antithesis of moral behavior. That AVs with PES would in most cases exemplify this antithesis is not just armchair speculation about human nature but something corroborated by empirical research. For example, in one poll, 64% of participants answered that they would even sacrifice a child in order to save themselves (Millar

2017. 25); other studies reveal that "[a]lthough people tend to agree that everyone would be better off if AVs were utilitarian (in the sense of minimizing the number of casualties on the road), these same people have a personal incentive to ride in AVs that will protect them at all costs" (Bonnefon et al. 2016. 1575). As a matter of fact, in order for it to fail by deontological standards, especially those set by Immanuel Kant (1785/1996), an AV with PES need not be *sensu stricto* selfish, that is, contributing exclusively to the well-being of its passenger. Just as unacceptable would be any other arbitrary or heteronomous motivation or reason – for example, positive or negative attitudes towards someone's race, sex, ethnicity or age – for distinguishing between traffic participants whose lives are worth saving from those whose lives are not worth saving.

PES also violates another important deontological principle: the prohibition against using persons "merely as means" (sometimes referred to as the "personhood" principle). In Kant's words, a human being "can never be used merely as a means by anyone (not even by God) without being at the same time himself an end" (1788/1996. 245). If I program my AV to systematically sacrifice anyone else in order to save my own life, this obviously amounts to using other persons merely as means. People treating each other as means, of course, is morally unproblematic as long as they do not treat each other *merely* as means, in the sense that everyone involved either explicitly agrees to a specific scheme of (inter) action or that their consent can be reasonably presumed (O'Neill 1994. 44). For example, I use the delivery driver as a means to get my pizza and he uses me as a means to earn his wages. The problem appears when people are treated *merely* as means and would not consent to such treatment if they were asked. For example: A and B survived a plane crash on a desert island. A kills B in his sleep, so he can eat him and survive until the rescuers arrive. B did not consent – and probably would not if A asked him – to be used in this way. PES is structurally similar and, for this reason, similarly problematic. One cannot reasonably presume that any person – in the other vehicle, or on the sidewalk or crosswalk – has consented to be killed (to be used merely as a means), so that I can continue living. I can reasonably presume my delivery driver's consent to be used as a means to get my pizza, but I cannot presume his consent to be run over by my AV to stop it from crashing into the back of a truck.

Partiality is one of the clearest utilitarian deficits of PES. Utilitarians insist on "the greatest happiness for the greatest number", but they also insist that this happiness is achieved in an impartial way. In John Stuart Mill's formulation: "[T]he happiness which forms the utilitarian standard of what is right in conduct is not the agent's own happiness, but that of all concerned" and "between his own happiness and that of others, utilitarianism requires him to be as strictly impartial as a disinterested and benevolent spectator" (1863/1998. 64). Peter Singer uses the "scales" metaphor: "True scales favour the side where the interest is stronger or where several interests combine to outweigh a smaller

number of similar interests, but they take no account of whose interests they are weighing" (2011. 20–21). An AV with PES that prioritizes its passengers' lives and interests over all other lives and interests – an option, as we have seen, that would be adopted by the majority of AV passengers – would obviously violate this utilitarian requirement of strict impartiality and disinterested benevolence.

A more serious utilitarian deficit of PES is its strong tendency – in comparison to other types of ethics settings – to bring about the worst possible consequences. If most AVs are set to protect their passengers' lives at all costs, including the cost of sacrificing any number of other lives, that should unquestionably, in the long run, increase the total number of traffic fatalities. This outcome is diametrically opposed to the fundamental utilitarian (consequentialist) principle of minimizing suffering and maximizing happiness for the greatest number of people possible. An argument to the same effect, presented in game-theoretical terms, is offered by Gogoll and Müller (2017). They maintain that allowing people to personally choose their own ethics settings would create "prisoner dilemma" circumstances in which everyone's probability of dying in traffic increases. Their basic point is this: even individuals disposed to choose "moral" PES (sacrificing themselves to save the greater number of others), as opposed to "selfish" PES (sacrificing any number of others to save themselves), would at some point realize that they are taken advantage of by selfish individuals. In this kind of environment, guided by rationality and in pursuit of their own interest, they would eventually switch to "selfish" ethics settings themselves, contributing thus to the creation of "a world in which nobody is ready to sacrifice themselves for the greater number" and "the number of actual traffic casualties is necessarily higher" (Gogoll–Müller 2017. 694). The proposed solution to this dilemma – to be analyzed in the next section – is MES:

> This leaves us with the classical solution to collective action problems: governmental intervention. The only way to achieve the moral equilibrium is state regulation. In particular, the government would need to prescribe a mandatory ethics setting (MES) for automated cars. The easiest way to implement a MES that maximizes traffic safety would be to introduce a new industry standard for automated cars that binds manufacturers directly. The normative content of the MES, that we arrived at through a contractarian thought experiment, can easily be summarized in one maxim: *Minimize the harm for all people affected!* (Gogoll–Müller 2017. 695)

IV. THE CASE AGAINST MES

The deontological deficits of MES are practically the mirror image of the deontological deficits of PES: whereas the major problem with PES is not autonomy but selfishness, the major problem of MES is not selfishness but autonomy.

As the German report on *Automated and Connected Driving* correctly recognizes, MES implies that "humans would, in existential life-or-death situations, no longer be autonomous but heteronomous" and that the state would act "in a very paternalistic manner and prescribing a 'correct' ethical course of action" (BMVI 2017. 16). MES would basically suspend an individual's capacity for ethical decision-making in situations – those with human lives at stake – in which the exercise of this capacity might be most needed. In other words, autonomous decision-making and moral agency would be substituted by algorithmic ("heteronomous") decision-making and preprogrammed agency. Since the specifics of this decision-making, by the definition of MES, would be prescribed and enforced by the state, it may actually be inadequate to talk about it as *moral* or *ethical* decision-making – in the same way as it would be erroneous to talk about any state prescribed and enforced norms as moral or ethical. In short, deontologists could claim that MES, as a consequence of its suspension of individual autonomy and moral agency, is actually a negation of ethics and should not be classified as an "*ethics* setting" at all.

An equally important deontological deficit of MES is that it implies using persons merely as means, in the sense of sanctioning a practice of sacrificing some persons – when traffic circumstances dictate it – to save the greater number of others. The fact that this would not be done by other persons (as was the case with PES), but by the state, is morally irrelevant. If a human being, as Kant said, "can never be used merely as a means by anyone (not even by God)", then they cannot be used merely as a means even by the state. The German report similarly points out that "offsetting of victims" by AVs is impermissible because "the individual is to be regarded as 'sacrosanct'" and "equipped with special dignity" (BMVI 2017. 18–19). It is important to notice that the wrongness of using persons merely as means here does not essentially stem from the fact that it would be performed by machines (which is a common ethical objection to many similar uses of AI systems). It would be wrong even if it was performed by human beings. Imagine that a time machine is invented that allows humans, at any given moment, to "freeze" time and everything that happens. They can "freeze" dilemmatic situations with AVs before they play out and allow human experts – some kind of a time travelling ethics committee – enough time to decide how to resolve them (for example, whether to sacrifice pedestrians or passengers). Assuming that persons affected by these decisions would not be consulted, the time travelling ethics committee would be treating them merely as means in the same way that MES would.

A possible reply to "autonomy" and "personhood" objections is that their force diminishes if all or the majority of citizens decide, through some kind of democratic procedure, that they wish to trade parts of each individual's autonomy and personhood for the reduction of everyone's chances of being killed in traffic. The problem with this reply is well-known from ethical debates on a

variety of sensitive issues like abortion, euthanasia or capital punishment: the majority opinion is not necessarily the morally right opinion. A public referendum with any percentage of votes – tight votes especially – either approving or disapproving any of these practices does not settle the fundamental ethical question of their rightness or wrongness (except, maybe, for radical ethical relativists). As an institutional arrangement that will require almost daily choices between human lives, MES would surely become an extremely sensitive issue likely to split public opinion. However, in view of the diversity and value pluralism of contemporary democratic societies, it seems unsatisfactory to use any form of democratic decision-making as a tiebreaker for moral disputes with far-reaching consequences like the one over MES. For the same reason, is does not seem promising to use it to neutralize deontological objections as complex as autonomy or personhood.

The main problem with MES, from the deontological perspective, is the fact that it is a utilitarian scheme of action and all such schemes, in John Rawls's formulation, have to be rejected because they disregard "the distinction between persons" (1971/1999. 24). According to Rawls, it is impermissible "that the sacrifices imposed on a few are outweighed by the larger sum of advantages enjoyed by many" (1971/1999. 3) and, "under most conditions, at least in a reasonably advanced stage of civilization, the greatest sum of advantages is not attained in this way" (1971/1999. 23). Nevertheless, it might be too hasty to conclude that MES, despite its central goal of minimizing the harm for all people affected, would be mechanically taken on board by utilitarians. The way in which MES would accomplish this goal is likely to have harmful side effects that most utilitarians tend to invoke when they dismiss some other, in many respects similar, proposals. As an initial illustration, consider the following hypothetical example:

> You have five patients in the hospital who are dying, each in need of a separate organ. […] You can save all five if you take a single healthy person and remove his heart, lungs, kidneys, and so forth, to distribute to these five patients. Just such a healthy person is in room 306. He is in the hospital for routine tests. Having seen his test results, you know that he is perfectly healthy and of the right tissue compatibility. […] The other five patients can be saved only if the person in Room 306 is cut up and his organs distributed. In that case, there would be one dead but five saved. (Harman 1977. 3–4)

In terms of the number of lives to be saved, cutting up the person in room 306 seems to make perfect utilitarian sense: "one dead but five saved" sounds much better than "one saved but five dead". Most utilitarians, however, are more refined than that and numbers are not the only thing that matters in their moral reasoning. They tend to dismiss proposals like cutting up the person in room 306, because they believe that any similar practice, once it is allowed and be-

comes publicly known, is likely to have a series of harmful side effects. According to rule utilitarians like Richard Brandt (1965/2003), for example, a rule that allows one healthy person to be sacrificed in order to save five dying patients might, in the long run, bring about an even greater loss of lives (e.g. due to the growing distrust of doctors or the fear of visiting hospitals). An advocate of R. M. Hare's (1981) two-level utilitarianism could claim that doctors, due to their inherent human limitations and biases, would more often than not make wrong judgments about exactly who and when should be sacrificed, so that the greatest number of lives can be saved. Since they are very likely to have catastrophic consequences, calculations like these should not be allowed to be part of doctors' everyday work. Peter Singer's preference utilitarianism also leaves plenty of room for rejection of similar proposals and practices:

> If […] we decided to perform extremely painful or lethal scientific experiments on normal adult humans, kidnapped at random from public parks for this purpose, adults who entered parks would become fearful that they would be kidnapped. The resultant terror would be a form of suffering additional to the pain of the experiment. (Singer 2011. 51–52)

> If I am a person, I know that I have a future. I also know that my future existence could be cut short. If I think that this is likely to happen at any moment, my present existence will be fraught with anxiety and will presumably be less enjoyable than if I do not think I am likely to die for some time. If I know that people like myself are very rarely killed, I will worry less than if the opposite is the case. (Singer 2011. 77)

The utilitarian logic behind the examples mentioned so far can be captured as follows: Although an action may have some positive immediate effects (for example, five lives saved at the expense of one), there is an overriding reason against performing that action as long as it, once becoming publicly known, is likely to have negative side effects across the population at large and continuing indefinitely into the future (for example, the resultant terror, fear and anxiety at the individual and social level). Another useful illustration of this logic is the hypothetical example of "survival lottery" by John Harris (1975/1986). We are invited to imagine two dying patients, Y and Z, trying to persuade doctors to save their lives by acquiring healthy organs in a unique way:

> Y and Z put forward the following scheme: they propose that everyone be given a sort of lottery number. Whenever doctors have two or more dying patients who could be saved by transplants, and no suitable organs have come to hand through "natural" deaths, they can ask a central computer to supply a suitable donor. The computer will then pick the number of a suitable donor at random and he will be killed so that the lives of two or more others may be saved. (Harris 1975/1986. 89)

One possible reason for rejecting "the institution of the survival lottery", according to Harris (1975/1986. 92), is that its "harmful side effects in terms of terror and distress to victims, witnesses, and society generally" would be similar to the harmful side effects "occasioned by doctors simply snatching passers-by off the streets and disorganizing them for the benefit of the unfortunate." This "lottery scheme", as Harris emphasizes, "would eliminate the arbitrariness of leaving the life and death decisions to the doctors, and remove the possibility of such terrible power falling into the hands of any individuals, but the terror and distress would remain" (1975/1986. 92). In what follows, it will be argued that MES bears sufficient resemblance to actions like "cutting up the person in room 306", "kidnaping people at random from public parks for lethal experiments" and the "survival lottery" itself, to be rejected on the very same utilitarian grounds.

MES and "cutting up the person in room 306" are analogous due to the decisive role that randomness plays in them. Assume that the person in room 306 ends up being cut up and his organs distributed to five dying patients. It happened only because he, accidentally, visited a particular hospital on a particular day and was outnumbered by five dying patients that were, accidentally, in the same place at the same time and could be saved by his organs. Had he decided to visit another hospital (or the same hospital on another day), he would still be alive. MES would also sacrifice a person only because she, accidentally, crossed a particular street at a particular time and was outnumbered by several other people that were, purely by chance, in the same place at the same time and could be saved by sacrificing her. Had she decided to cross some other street (or the same street at a different time), she would still be alive. This kind of accidental factor or randomness, if allowed (and publicly announced) to influence life or death decisions in everyday circumstances, would undoubtedly cause enough "terror and distress to victims, witnesses, and society generally" to justify the utilitarian rejection of any similar scheme of action.

There is something more problematic with MES than with "cutting up the person in room 306". The conditions that have to be met, namely, for doctors to even begin considering the proposal of "cutting someone up" would be exceptional: What is the probability of (a) a healthy person (b) visiting the hospital for routine tests, (c) having his tissue compatible with five patients (d) each in need of a different organ, that are (e) already present in the hospital? This probability must be extremely low, but there is no doubt that most utilitarians would still reject any similar scheme of action – especially if it should become publicly known – as not worth the risk of harmful side effects. The problem with MES is that the probability of finding oneself in a dilemmatic situation, potentially as the person that has to be sacrificed by an AV, will be significantly higher. This much should be clear already from the fact that a large portion of the population participates in or is somehow affected by road traffic on a daily basis. Moreover, the probability of such an event will become even higher if AVs, as the Institute of

Electrical and Electronics Engineers (IEEE 2012) has predicted, "will account for up to 75 percent of cars on the road by the year 2040." What follows is that with state-wide implementation of MES, very few will be able to say that people like themselves "are very rarely killed", that their "future existence" is unlikely to be "cut short" and that they, therefore, have no reason to worry about MES.

As the final variation of the same utilitarian argument against MES, imagine MES 2.0 – an advanced version of MES that takes into account not only the number of people involved in dilemmatic situations (either as passengers or as pedestrians), but additional factors as well, such as their health status, age, profession, number of children and criminal record. The collection and use of such sensitive data – essentially in order to profile individuals for their suitability to be saved or sacrificed for the greater good – would be perceived by the general public as something negative and intimidating. Moreover, if its functioning will depend on technologies like machine learning or self-learning algorithms, it could be extremely difficult, from the technical point of view, to explain to the public how and on the basis of which data MES 2.0 makes its life or death decisions. A purely technical issue like this – also known as the "black box" problem of algorithms – would easily morph into a moral and political issue: any non-transparency, inexplicability or secrecy related to tools like MES, especially when they are controlled by state officials, tends to fuel suspicions and fear of things like corruption, discrimination or even totalitarianism. Bearing in mind, moreover, that AVs are "the first robots to be integrated with society at any significant scale" that might "set the tone for other social robotics, especially if things go wrong" (Lin–Jenkins–Abney 2017. ix), these suspicions and fear provide a solid utilitarian argument against MES.

It is possible to remain sceptical about the idea of MES as a cause of distress, anxiety and fear. Moreover, the opposite claim could be argued for: given that accidents that already happen with conventional vehicles do not trigger any systematic distress, anxiety and fear, AVs with MES could actually, by minimizing everyone's chances of being killed in traffic, prevent the occurrence of any similar distress, anxiety and fear. One problem with such a defense of MES is that practices like "cutting up the person in room 306" or the "survival lottery" could be justified in a similar way (by arguing, for example, that they would improve the chances of survival of all hospitalized persons or all members of society), but they would still be perceived as serious and morally unacceptable sources of distress, anxiety and fear. Another problem is that individuals might not care (although perhaps irrationally) about the statistical advantages of MES as much as they care about some other things it might interfere with, like the freedom to make their own decisions in life or death situations, a desire to protect their own lives or the lives of their family members first, or even – as we shall see in the next section – a commitment to certain moral principles and values. It should be emphasized, however, that the objective of this section was not to answer

the empirical question about the psychological effects of MES. Its objective was primarily conceptual: to identify similarities between the idea of MES and hypothetical scenarios that utilitarians themselves tend to reject and to show, in this way, that the moral damage potentially generated by MES need not be only deontological, but also utilitarian.

V. THE CASE FOR NES

Implementing either PES or MES, due to their unavoidable violation of several principles central to both deontology and utilitarianism, is bound to cause serious moral damage. In a nutshell, the deontological deficits of PES are the expected selfishness and using other persons merely as means, while its utilitarian deficits are the expected partiality and the tendency to bring about the worst possible outcomes in terms of the number of traffic fatalities. The deontological deficits of MES are the suspension of individual autonomy and using other persons (this time by the state) merely as means, while its crucial utilitarian deficit is the high potential to bring about harmful side effects – like distress, anxiety and fear at individual and social level – which most utilitarians anticipate and invoke when they reject some highly similar schemes of action. The presence of moral damage constituted by these moral deficits solves our initial trilemma: PES and MES have to be excluded and NES – that is, AVs unable to choose one human life over the other – remains the only plausible option.

It should be recognized that MES, thanks to its impartial distribution of harms and benefits among all those affected by its decisions, outcompetes both PES and NES in minimizing the number of traffic fatalities. However, a combination of two reasons, one statistical and the other one moral, is what makes NES the only plausible option. The statistical reason is that, when it comes to minimizing the number of traffic fatalities, NES outcompetes PES and is still, therefore, the second-best solution to how AVs should behave in dilemmatic situations. (Remember that PES has a practically inbuilt tendency to maximize the number of traffic fatalities whenever that saves the AV's passenger.) The moral reason should be familiar by now: NES causes no moral damage comparable to the one caused by either PES or MES. NES should be preferred to its alternatives, simply put, thanks to the best ratio of the expected success in minimizing the number of traffic fatalities to the expected range of its moral damage. In order to explicate this point further, consider the following hypothetical case by Bonnefon, Shariff and Rahwan:

> Say that two competing companies market self-driving cars that both eliminate 80% of fatalities, but one company's cars split the remaining fatalities equally between passengers and pedestrians, whereas the other company's cars split the remaining

> fatalities nine-to-one in favor of their passengers. Consumers would flock to the cars of the second company, and pedestrian risks would gradually inflate to unacceptably unfair levels. (Bonnefon at al. 2019. 504)

Under the assumption that AVs without any kind of ethics settings eliminate 80% of traffic fatalities and that AVs with some kind of ethics settings eliminate the remaining 20%, how can it be that this additional reduction of traffic fatalities does not suffice to compensate for the moral damage that any ethics settings might cause? How can saving moral principles or abstract values be more important than saving human lives? One answer to these questions could be hiding in the hypothetical case itself: If using PES to eliminate the additional 20% of traffic fatalities is considered unacceptably unfair to pedestrians as a *group*, then using MES to achieve the same 20% improvement should be considered unacceptably unfair to any *individual* (passenger or pedestrian) killed by an AV only because she happened to be (from her perspective) in the wrong place at the wrong time (although in the right place and the right time from the perspective of those saved by sacrificing her life). Illustrated by analogy with doctors cutting up one person as a "donor" and distributing his organs to five dying patients: It would surely be unfair to select this person from a specific group of potential donors (for example, already hospitalized patients, persons over 50 or people without children), but it does not seem any fairer to select this person at random from visitors of public parks, people on the street or – for that matter – the general population.

Another answer is more general: to save as many lives as possible is desirable, but the way they are saved is not morally irrelevant and it may, depending on the situation, constitute a reason against saving them. Consider negotiating with terrorists, torturing kidnappers, paying ransoms, collective punishment, wiretapping of ordinary citizens, buying and selling of newborns for adoption, etc. Although practices like these, in certain circumstances, could save lives, they tend to be widely rejected as morally unacceptable. This rejection is typically defended in either deontological or utilitarian terms, by claiming that allowing such practices violates basic human rights or that it sets dangerous precedents with harmful side effects. It is interesting, moreover, that some of these practices are considered morally unacceptable even in emergency situations like war. It is particularly interesting that there are numerous voices, among both scholars and the general public, opposed to any wartime use of military robots or autonomous weapons. One frequently mentioned reason for this opposition is that these weapons could not distinguish combatants as legitimate targets from innocent civilians as illegitimate ones. The lesson for the ethics settings debate, at the very least, is the following: if unintentionally sacrificing innocent lives is a serious reason to reject autonomous weapons in extraordinary situations such as war, it is too unrealistic to expect any serious acceptance of AVs programmed to intentionally choose one innocent human life over the other in ordinary situations like daily traffic.

VI. CONCLUSION

The primary purpose of this article was not to decide which ethical position, deontology or utilitarianism, provides a more fertile ground for building a case against AVs with ethics settings. Its primary purpose was to argue that any type of ethics settings capable of choosing one human life over the other is bound to cause serious moral damage resulting from the violation of several principles central to both deontology and utilitarianism. AVs without ethics settings are the preferred solution because that option sufficiently minimizes the number of traffic fatalities without causing any comparable moral damage. The overall conclusion of the article is that AVs with ethics settings will remain not only a bridge that we should not cross, but most likely a bridge that most people will never have a serious intention of crossing.

REFERENCES

Awad, Edmond – Sohan Dsouza – Richard Kim – Jonathan Schulz – Joseph Henrich – Azim Shariff – Jean-François Bonnefon – Iyad Rahwan 2018. The Moral Machine Experiment. *Nature*. 563. 59–64.

BMVI 2017. *Automated and Connected Driving*. Bundesministerium für Verkehr und digitale Infrastruktur.
www.bmvi.de/SharedDocs/EN/publications/report-ethics-commission.pdf?__blob=publicationFile

Bonnefon, Jean-François – Azim Shariff – Iyad Rahwan 2016. The Social Dilemma of Autonomous Vehicles. *Science*. 352(6293). 1573–1576.

Bonnefon, Jean-François – Azim Shariff – Iyad Rahwan 2019. The Trolley, the Bull Bar, and Why Engineers Should Care about the Ethics of Autonomous Cars. *Proceedings of the IEEE*. 107(3). 502–504.

Brandt, Richard B. 2003 [1965]. Toward a Credible Form of Utilitarianism. In Stephen Darwall (ed.) *Consequentialism*. Oxford, Blackwell. 207–235.

Foot, Philippa 1967. The Problem of Abortion and the Doctrine of Double Effect. *Oxford Review*. 5. 5–15.

Gogoll, Jan – Julian F. Müller 2017. Autonomous Cars: In Favor of a Mandatory Ethics Setting. *Science and Engineering Ethics*. 23(3). 681–700.

Hare, Richard M. 1981. *Moral Thinking: Its Levels, Method, and Point*. Oxford, Oxford University Press.

Harman, Gilbert 1977. *The Nature of Morality: An Introduction to Ethics*. New York, Oxford University Press.

Harris, John 1986 [1975]. The Survival Lottery. In Peter Singer (ed.) *Applied Ethics*. New York, Oxford University Press. 87–95.

IEEE 2012. Look Ma, No Hands! Institute of Electrical and Electronics Engineers. https://www.ieee.org/about/news/2012/5september-2-2012.html

Kant, Immanuel 1785/1996. *Groundwork of The Metaphysics of Morals*. In Immanuel Kant: *Practical Philosophy*. Transl. by Mary J. Gregor. Cambridge, Cambridge University Press.

Kant, Immanuel 1788/1996. *Critique of Practical Reason*. In Immanuel Kant: *Practical Philosophy*. Transl. by Mary J. Gregor. Cambridge, Cambridge University Press.

Lin, Patrick – Ryan Jenkins – Keith Abney 2017. Preface. In Patrick Lin – Ryan Jenkins – Keith Abney (eds.) *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*. New York/NY, Oxford University Press. ix–xiii.

Mill, John Stuart 1863/1998. *Utilitarianism*. Oxford, Oxford University Press.

Millar, Jason 2017. Ethics Settings for Autonomous Vehicles. In Patrick Lin – Ryan Jenkins – Keith Abney (eds.) *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*. New York, Oxford University Press. 20–34.

Nyholm, Sven 2018a. The Ethics of Crashes with Self-Driving Cars: A Roadmap, I. *Philosophy Compass*. 13(7). https://doi.org/10.1111/phc3.12507

Nyholm, Sven 2018b. The Ethics of Crashes with Self-Driving Cars: A Roadmap, II. *Philosophy Compass*. 13(7). https://doi.org/10.1111/phc3.12506

O'Neill, Onora 1994. A Simplified Account of Kant's Ethics. In James E. White (ed.) *Contemporary Moral Problems*. St. Paul, West Publishing Company. 43–48.

Rawls, John 1971/1999. *A Theory of Justice. Revised Edition*. Cambridge, Harvard University Press.

Singer, Peter 2011. *Practical Ethics*. 3rd edition. Cambridge, Cambridge University Press.

Thomson, Judith Jarvis 1976. Killing, Letting Die, and the Trolley Problem. *Monist*. 59. 204–217.

Miklós Hoffmann

Science as a Human Vocation and the Limitations of AI-Based Scientific Discovery*

Abstract

In his essay *Science as a Vocation*, Max Weber took the essence of scientific activities to consist in specialisation and enthusiasm. His arguments, together with works by Michael Polanyi (Mihály Polányi) and others, are explored and compared with recent results and expectations of automatised, artificial-intelligence-driven scientific discovery. Our aim is to show that artificial intelligence systems (AI systems) – while they can evidently and effectively support everyday scientific activities as useful tools – are not, in themselves, able to produce genuine invention, are not suitable for breakthrough scientific discovery. And this limitation, we argue, is due to AI systems' inability for specialisation and their lack of enthusiasm. Our observation is that while selection by intrinsic interest is unavoidable and an essential part of science, this interest is unquantifiable and unmetrisable by an objective function, therefore cannot be learned by an AI system. We conclude that being a scientist full of passion and with the ability of selection remains humans' intellectual privilege.

Keywords: scientific discovery, artificial intelligence, invention, enthusiasm

I. INTRODUCTION

In February 1996 the reigning world chess champion Garry Kasparov was defeated by the IBM computer Deep Blue. Although Deep Blue controlled the white pieces, and right after that game Kasparov won the next game, and was the overall winner of the six-game chess match (Kasparov versus Deep Blue: 4–2), this date still marks an unusual technological achievement: this was the first win of an artificial-intelligence-driven system (AI system) over the highest-ranked human specialist in a specific field of expertise. One year later Deep Blue defeated Kasparov 3.5–2.5 in another six-game chess match.

* This research was supported by the grant EFOP-3.6.1-16-2016-00001 (Complex improvement of research capacities and services at Eszterházy Károly University). The author would like to thank the anonymous reviewers for their thoughtful comments on this work.

Almost exactly 20 years later, in March 2016, Lee Sedol, one of the greatest players of Go, a highly complex strategy board game popular in East-Asia, was defeated by Google's AlphaGo software. Go has far more variations than chess, and strategies are more complicated (Bouzy 2001), therefore this win is another important milestone in the development of artificial intelligence. The event was selected as one of the scientific breakthroughs of the year 2016 by Science Magazine.[1] And the momentum was unstoppable: in less than one year, AI software defeated over 100% of the best poker players in several poker tournaments. Why are these developments so interesting? While chess and Go are called games with complete information – that is, players possess full information about their opponents and their potential (straightforward or surprising) actions – poker is clearly a game with incomplete information. The possibility of bluffing makes a poker game somewhat independent of the consequences of previous steps, it liberates players from the restrictions of logic, therefore opponents need to study not only combinations and strategies – beside learning game rules, computers must also learn the behaviour and attitude of other players (Moravčík 2017).

Besides board and card games, it was a natural next step to compare humans and AI systems in other fields as well. In this paper we intend to study how the rapid development of AI may impact on human scientific activity, science as a vocation: more specifically, the potential automatisation and algorithmisation of scientific discovery.[2] Some are sceptical about such impact. Others are warning the sceptic: it may be prudent to reassess doubts given that Kasparov and Sedol were also antecedently doubtful as to whether an AI system could beat them. For example, Nobel laureate scientist Wilczek (2016) (among others) strongly believes that scientific discovery will soon be fully automatised. According to Wilczek and other scientists (see e.g. Kitano 2016) it is a realistic scenario that an AI system will be the best physicist and will be able to win the Nobel Prize in the near future.

Attempts to find scientific achievements through automated discovery have a long history during which tools and concepts have significantly evolved. In this paper the label "AI system" is taken to encompass earlier serial-computation approaches as well as more recent machine learning approaches, genetic algorithm based methods, or their fusion (for an overview of these methods and their history see Alai 2004). We will call all these AI systems without differentiating among them, because we believe that there are two significant common attributes to all of these methods: their data-driven approach and their algorithmic, procedural nature. Regardless of the method and tool, artificial intelligence requires data – a large amount of training data in which it can find typical patterns

[1] Besides gravitational waves and customised proteins, see *Science* 354. 1518–1523.

[2] Throughout the paper notions like "algorithm" and "computer" are used in the usual manner for software and hardware tools developed for determined computation executed in a finite number of steps. However, we note here that illuminating discussion about these notions is now under way in the literature, see e.g. Rapaport 2018.

and correlations. And this is done by a procedure, an algorithm, even in the case of the most sophisticated neural network methods.

One may ask whether scientific discovery or even the scientific description of the world can have a substantially different path than what we have experienced throughout the history of science. We cannot answer this question here, but the fact remains that no alternative approach has been envisioned so far: all the attempts at automatised scientific discovery follow our classical path and a potential new, uncharted path may well diverge considerably from what we now call science and knowledge. Nevertheless, our discussion remains in the classical framework: we consider scientific discovery and science as an enterprise whose results were, over many centuries, produced by human scientists.[3]

In this paper we intend to point out those substantial aspects of scientific discovery that make the personal involvement of human scientists inevitable, and consequently make the replacement of scientists by computer algorithms and artificial intelligence in the scientific process highly doubtful. Our arguments will extensively rely on Max Weber's stance, who saw the essence of scientific activities in specialisation and enthusiasm (Weber 1946). These key notions will be analysed in our study from the perspective of AI-driven scientific discovery. We aim to show that AI systems – while they can evidently and effectively support everyday scientific activities as useful tools – are not, in themselves, able to produce genuine invention, are not suitable for breakthrough scientific discovery. And this limitation, we argue, is due to AI systems' inability for specialisation and their lack of enthusiasm.[4]

One may think that specialisation cannot be an obstacle to AI in terms of automatised scientific discovery: for the computational and learning capacity of these algorithms can easily be focused on an arbitrary narrow field. However, as we will show, from a theoretical point of view, the specialisation requirement yields an insurmountable problem for artificial intelligence. Enthusiasm, as we will also point out, raises an even more difficult issue.

We put special emphasis on the enthusiasm-filled moment that anticipates scientific work. Max Weber writes about this moment as follows:

> Yet it is a fact that no amount of such enthusiasm, however sincere and profound it may be, can compel a problem to yield scientific results. Certainly enthusiasm is a prerequisite of the "inspiration" which is decisive. Nowadays in circles of youth there is a widespread notion that science has become a problem in calculation, fabricated

[3] This view also gives credibility to the thoughts of scientists from past centuries about science and scientific discovery, even if automatised scientific discovery was not an issue, or it was technically less developed in their time.

[4] From a Kuhnian perspective: artificial intelligence is able to support "normal science" through day-to-day experimental studies, but it cannot discover results forcing a paradigm shift.

> in laboratories or statistical filing systems just as "in a factory", a calculation involving only the cool intellect and not one's "heart and soul". First of all one must say that such comments lack all clarity about what goes on in a factory or in a laboratory. In both some idea has to occur to someone's mind, and it has to be a correct idea, if one is to accomplish anything worthwhile. And such intuition cannot be forced. It has nothing to do with any cold calculation. (Weber 1946. 135)

This – in our view, essential – moment, the birth of the first idea, the exciting promise of the discovery, the moment of entering the force field of the problem, I will call – applying a physical metaphor – *the gravity of invention*.

II. AI-DRIVEN SCIENTIFIC DISCOVERY – INABILITY FOR SPECIALISATION

The first research result about an AI system engaging in scientific discovery was published by Pat Langley and his colleagues (Langley et al. 1987). In this study an AI system was programmed by the research team to explore new scientific results based on a data set. In their groundbreaking study the most interesting aspect is the history-oriented approach, which, to some extent, already predisposes it towards verifying a preconceived outcome: during the training period, data fed into the AI system was selected from a certain historical period of science. Physical and chemical observations and laws known around the 17th and 18th centuries were learned by the system. Based on these data, the AI system "discovered" now well-known, but at-the-time new scientific results such as Ohm's law, Kepler's third law of planetary motion, and various chemical reactions.

However, besides these apparently successful outcomes the computer also "discovered" superseded scientific theories such as the phlogiston theory mistakenly put forth to explain oxidation. Moreover, other outcomes were true but totally uninteresting from a scientific point of view. Note here that those results, such as Kepler's law, discovered by the AI system, can be deduced (and in fact have been subsequently discovered by Kepler) by systematically tracking the available observational data over a long period of time. In other words, systematic computational work on observational data can readily lead us to this discovery. The phrase "systematic" is used here as the opposite of "heuristic", following a distinction drawn by Michael Polanyi:

> The difference between the two kinds of problem solving, the systematic and the heuristic, reappears in the fact that while a systematic operation is a wholly deliberate act, a heuristic process is a combination of active and passive stages. A deliberate heuristic activity is performed during the stage of Preparation. If this is followed by a period of Incubation, nothing is done and nothing happens on the level of consciousness during this

> time. The advent of a happy thought (whether following immediately from Preparation or only after an interval of Incubation) is the fruit of the investigator's earlier efforts, but not in itself an action on his part; it just happens to him. And again, the testing of the "happy thought" by a former process of Verification is another deliberate action of the investigator. Even so, the decisive act of discovery must have occurred before this, at the moment when the happy thought emerged. (Polanyi 1974. 134)

A scientific discovery is called systematic if the final result is reached by a series of intentional, algorithm-based steps, even if these steps are very complicated. By contrast, the discovery is heuristic if – beside the above mentioned steps – it is based on one or more unanticipated, unenforceable moments, which cannot be explained as a simple logical consequence of preceding steps. These are the moments of Weberian inspiration, the moment when a – perhaps brilliant – thought arises. For example, contrary to Kepler's law, the thought of the heliocentric system by Copernicus cannot be the outcome of a systematic discovery, since observational data available given that era's level of accuracy provided stronger support for the Ptolemaic system. Analogously, the theory of general relativity by Einstein cannot be algorithmically derived from the observational data of that age – it was experimentally proven only decades after the publication. Since every result the AI system can produce is inherently based on the analysis of available observational data, it can yield systematic scientific discovery, but we claim that brilliant heuristic moments and thoughts lie outside the repertoire of an AI system.

One may think that even if we cannot expect from AI systems groundbreaking discoveries in the natural sciences or mathematics (discoveries that require the power of a compelling paradigm change), many useful and interesting results in a specialised narrow subfield may still be gleaned by an AI system. And this leads us to the question of specialisation, whose importance was also emphasised by Weber. But specialisation certainly involves selection: scientists have to select among topics, within the given topic they have to select among related theorems, laws, data which are to be learned, improved or further developed. Moreover, one even has to select among the potentially solvable problems and among the provable theorems. Selection is unavoidable due to our limited resources, but there is an even more important aspect: the intrinsic interest of the problem. It is worth citing Michael Polanyi again on this:

> An affirmation will be acceptable as part of science, and will be the more valuable to science, the more it possesses: (1) certainty (accuracy) (2) systematic relevance (profundity) (3) intrinsic interest. (Polanyi 1974. 143)

While (1) and (2) sound natural requirements in the realm of scientific inquiry, (3) is a property that is difficult to make precise, yet it is of central importance. We clearly have no exact tools or algorithms or conditions to evaluate effectively

the level of interest of a scientific statement. No one can assess based on exact criteria what theorem or law is more interesting (or will be in the future) than another statement of physics, chemistry or mathematics. Having said that, selection by intrinsic interest looks not only unavoidable, but also essential. It is evident that our (human or articifical) intellectual capacity is restricted in terms of time and computational power, therefore it is highly beneficial to focus this capacity on problems which may yield higher "gains", and can improve our scientific knowledge in a more effective way. The higher the intrinsic interest of a problem, the stronger its gravity of invention. Stronger gravity can also affect, influence more scientists. We provide some examples for such an interest arising among mathematicians because – compared to the natural sciences – mathematics is a field where scientists can formulate new valid statements in a relatively easy way, thus in relatively large numbers.

Since mathematics is a cumulative, aggregate field of science, whenever a statement is correctly proved, it will be part of mathematics forever. The so-called Ulam's dilemma (Ulam 1976) describes the ever-more-complex situation as follows: in mathematics (and partly in theoretical physics) we have discovered so many theorems, and scientists extend this list daily by such a vast amount of valid statements, that nobody is able to overview the entire field, only some sufficiently small subfield.[5] The only solution to this dilemma is specialisation, also encouraged by Weber. Specialisation means selection: selection among theorems, among subfields, among problems. This selection, however, is not a drawback, not a restriction, not a systemic limitation, contrary to how one may view it at first glance. Selection is the essence of scientific discovery. It is worth citing here one of the greatest mathematicians of the 19th and 20th centuries, Henri Poincaré:[6]

> What, in fact, is mathematical discovery? It does not consist in making new combinations with mathematical entities that are already known. That can be done by anyone [*even a computer – M.H.*], and the combinations that could be so formed would be infinite in number, and the greater part of them would be absolutely devoid of interest. Discovery consists precisely in not constructing useless combination, but in constructing those that are useful, which are an infinitely small minority. Discovery is discerment, selection. (Poincaré 2009. 50)

[5] In his book, Stanislaw Ulam estimated the number of yearly published mathematical theorems around 200 000 – and this number evidently further increased (probably exponentially) in recent decades.

[6] In the original version: "Qu'est-ce, en effet, que l'invention mathématique? Elle ne consiste pas à faire de nouvelles combinaisons avec des êtres mathématiques déjà connus. Cela, n'importe qui pourrait le faire, mais les combinaisons que l'on pourrait former ainsi seraient en nombre infini, et le plus grand nombre serait absolument dépourvu d'intérêt. Inventer, cela consiste précisément à ne pas construire les combinaisons inutiles et à construire celles qui sont utiles et qui ne sont qu'une intime minorité. Inventer, c'est discerner, c'est choisir." (Poincaré 1912. 48)

Invention thus practically amounts to selection when done well, and well in time. But such selection cannot be algorithmisable, since it is not a mechanically scientific, but rather a meta-mathematical selection. If we start from an axiomatic system, say, the Peano-axioms of natural numbers, then human as well as artificial intelligence can prove many valid statements, for example, that there is an infinite number of primes; and can falsify many other untrue statements, such as there is no even prime. Moreover, artificial intelligence can evidently "produce" many more valid theorems and can falsify many more untrue statements in a given period of time than human scientists can. However, as Karl Popper[7] (1950) also points out, computers have no instruments or algorithms to draw a distinction between what are – in our view – interesting, thought-provoking, ingenious statements and statements which are totally uninteresting (although true). A very simple, yet convincing example of Popper's can further illuminate this problem and make it plausible: besides the statement $2 + 1 = 3$, a computer will find infinitely many statements like $2 + 1 \neq 4$; $2 + 1 \neq 5$… and further statements like $2 + 1 \neq 3 + 1$; $2 + 1 \neq 4 + 1$, all arrived at based on the same set of starting axioms. For each substantial, interesting statement an AI system systematically generates infinitely many uninteresting yet valid statements.[8] Overall, the probability of observing the few promising ideas worth further investigation among the many-many uninteresting statements by the computer is very close to zero.

III. AI-DRIVEN SCIENTIFIC DISCOVERY – LACK OF ENTHUSIASM

As we have already mentioned, besides the ability, instinct and delight of specialisation, Max Weber has seen the substance of scientific activities in enthusiasm. What does enthusiasm – or lack thereof – mean in terms of science as a vocation? When engagement with a problem is externally driven (a typical example for most of us is solving a task provided by the teacher in a mathematics class) then one can feel the sense of duty or competition, the wish to surmount the hurdles related to the problem, but the extrinsic nature of motivation deprives us of feeling passion and enthusiasm. By contrast, if the motivation for solving the problem comes from an intrinsic interest, if an unforced and unforce-

[7] "A calculator may be able, for example, to produce proofs of mathematical theorems. It may distinguish theorems from nontheorems, true statements from false statements. But it will not distinguish ingenious proofs and interesting theorems from dull and uninteresting ones. It will thus 'know' too much — far too much that is without any interest." (Popper 1950. 194)

[8] Although it is not well defined what we mean by "interesting" and "uninteresting" results, mathematicians have a surprisingly well-functioning common intuition in judging the value of propositions. Overall this leads to the question of (un)metrisability of scientific interest, which we will discuss in the last section.

able seed of idea emerged in our head, then we will engage this problem with personal commitment, passion and enthusiasm.

In his famous book *Proofs and Refutations*, Imre Lakatos (1976) studied and demonstrated through several examples how a (mathematical) problem and invention may arise, among which here we briefly refer to one typical scenario.

Suppose that – as a beginner in maths – we study divisibility of numbers and we observe that every number whose last digit is 2 (such as 12, 22, 32 etc.) is divisible by 2. Meanwhile numbers whose last digit is 3 are not always divisible by 3 (for example 63 is divisible by 3, but 13 is not divisible). We find it interesting that there are numbers analogous to 2, for example numbers whose last digit(s) is 5 (or 10 or 25) are always divisible by 5 (or 10 or 25), call these last-divisible numbers. Meanwhile we find several numbers in the other class as well: for example 24 is divisible by 4, but 14 is not. Now we are right in the middle of the field of gravity of the problem, the gravity of invention, and the data we collected (last-divisible examples are 2, 5, 10, 25, 50, 100…) make the heuristic idea clear: all the last-divisible numbers are products of powers of 2 and 5, possibly including powers with exponent 0 (note, however, that not all numbers that are such products are last-divisible, for example, 5^4 isn't while 5^3 is). This is far from a rigorous proof, but the rest is simply a mechanical computation for formulating and justifying the precise statement.[9]

In the example described above and in many other examples Lakatos has presented (in a much more detailed form in his book), he describes the athmosphere of raising a problem and finding a heuristic solution. Here we intend to focus on one important aspect of this process: assuming an underlying principle in the collected examples and counterexamples, based on which one can heuristically create a conjecture is of utmost importance. The key moment is the perception of the first couple of aspects of the pattern, the excitement of foreseeing the potential existence of some (ir)regularity. This excitement is not about the foreseen result, but about the promise of an interesting result. It is the gravity of invention, the engagement of the scientist in the field of gravity of the problem. The first perception about the number 2 is not specifically exciting, but the moment of understanding that another number (3) works differently than 2 may put our mathematical thinking in action. Anticipating the promise of success, we try to find new examples and counterexamples. Finding these data it can happen that the problem turns out to be too simple, too trivial, or uninteresting. But it can also happen that the intrinsic interest of the problem drives us into a new field and activates our heuristic problem-solving abilities.

[9] The theorem in its final form is as follows: a number n is last-divisible if and only if $n = 2^p 5^q n = 2^p 5^q$, where $p, q \geq 0 p, q \geq 0$ and $0 \leq q - p + 1 \leq 40 \leq q - p + 1 \leq 4$.

Note that the first moment, the promise of a future fruitful cogitation is, in fact, not part of the heuristic problem-solving process. It is a "preheuristic" flash, yet it is essential in terms of mathematical discovery and in general in scientific vocation.

The start of gravity of invention, the passion of thinking, the way how the scientist is getting engaged by the problem, is unexplainable, unenforceable, and, more importantly, unpredictable. It cannot be foreseen, cannot be measured (as we will see soon in some detail). And overall this foreshadows that an AI system, evidently driven by external forces (i.e. programmers) when studying chess, making stock market transactions or proving geometric theorems, cannot be programmed for this central passion, for the enthusiasm towards science. Some aspects of this discussion ultimately lead us to the most fundamental questions of artificial intelligence, notably issues having to do with what it takes for a system to have mental states, and more specifically, mental states of the sort that can underpin goals, motivation. These questions are beyond the scope of this paper, but even if we suppose that future AI systems can have mental states, these states (and the change of them) are outcomes of a causal process, externally driven (by the programmer and partly by the input data). Therefore it is entirely unclear if and how enthusiasm and passion of thinking is achievable by artifical intelligence systems.

There is no doubt that computers, without any passion or enthusiasm, can indeed find interesting results in some fields of science. For example, the AI-driven computer called "Eve" has been searching and finding effective pharmaceutical components, carrying out a vast number of trials (see King 2018). A further recent example for this type of discovery is from the field of material science, reported by Tshitoyan (2019) and Kauwe (2020). AI systems can discover new materials or new properties of old materials, but only following the typical patterns of an extremely large training data set of information. However, if we are seeking to discover something atypical, a kind of material which achieves its extraordinary properties by leveraging a new mechanism that is not common in the training data set, it will be unlikely to identify it through AI-driven discovery (c.f. Kauwe 2020). Atypical discovery always needs heuristic impulse and vision. The following sentences from Michael Polanyi clearly demarcate the barriers:

> The heuristic impulse links our appreciation of scientific value to a vision of reality, which serves as a guide to enquiry. Heuristic passion is also the mainspring of originality. (Polanyi 1974. 169)

Automatic scientific discovery without enthusiasm can only happen through a great number of trials, through following or finding typical patterns. Without heuristic impulse there is no chance of realising and evaluating the potential value of a future discovery which may come from a certain direction of research.

Even if a significant discovery is found by an AI system, it is not necessarily able to realise its importance.

The scientist is not cold and unemotional during research, not even at the beginning, at the preheuristic moment of involvement. As, following Heidegger and Gadamer, István Fehér M. (2017. 15) formulated this succintly, the scientist always has a – positively considered – prejudice:[10]

> …with regard to the type of interpretation that is directed at texts, in most cases it is illusory to refer to what "stands there" in the text as decisive evidence. For what is first and foremost "there" – provided there is any sense in speaking of "standing there" – is not so much the text itself, but rather "the self-evident, undisputed preliminary prejudice [*Vormeinung*] of the interpreter".

We can extend this approach to non-textual (visual or machine-based) sources of scientific information as well: there is no decisive, original meaning of pictures, figures, graphs, equations, data, and software output. What exists is a meaning interpreted by the scientist who studies that source, and this meaning is filtered and fertilised through the preliminary prejudice [*Vormeinung*], and positively considered preconception [*Vorurteil*] of the scientist. An AI system does not possess and cannot be equipped with such a preconception and prejudice: computers can treat only the information "standing there" technically or syntatically without any relationship to the source of information, without any preliminary opinion, because these are all beyond (or rather before) the pure binary information. It is as yet entirely unclear how to equip an AI system with more about the subject matter of investigation than the pure binary data we have provided. Let's compare this to Polanyi's words:

> To see a problem is to see something hidden that may yet be accessible. The knowledge of a problem is, therefore, like the knowing of unspecifiables, a knowing of more than you can tell. But our awareness of unspecifiable things, whether of particulars or of the coherence of particulars, is intensified here to an exciting intimation of their hidden presence. It is an engrossing possession of incipient knowledge which passionately strives to validate itself. Such is the heuristic power of a problem. (Polanyi 1961. 466)

Note that knowing the problem is not the same as knowing the solution to the problem. To know problems, or even to feel problems requires the recognition of their hidden presence, and the excitement of this recognition, the possibility of invention is gravitating us to the search for a solution to the problem and to the application of heuristics. The latter, that is to say, our attempt to solve

[10] In this paragraph the author refers to Heidegger (1962. 141).

the problem, may or may not succeed, but the gravitational attraction already mentioned will trigger the process. Mathematics uses one concise word for all of this: conjecture. The mathematical conjecture is preconceived knowledge, prejudice, similar to what is discussed in Gadamer's legal example (Gadamer 2004. 194) as a preconceived judgement: something I think about the thing before I know the thing, which can be verified or falsified by subsequent careful examination.

The AI system tries to grasp the problem without prejudice, with the question "what is that?," while we begin to engage the problem because it already means something to us, so our question (according to Nietzsche) is: "what is that for me?".[11] With the question "what is that?" the computer searches for an objective constitution, an absolute meaning in all data. When evidently expecting two computers to analyze the same data to produce the same result, we actually discover and demonstrate limitations to artificial intelligence. Let's quote Gadamer again:

> The paradox that is true of all traditionary material, namely of being one and the same and yet of being different, proves that all interpretation is, in fact, speculative. Hence hermeneutics has to see through the dogmatism of a "meaning-in-itself" in exactly the same way critical philosophy has seen through the dogmatism of experience. (Gadamer 2004. 507)

The question "what is that for me?" can be answered differently even if two of us look at the same text, same data, same picture. Moreover, here, in addition to our semantic relation, the expression "means something to me" as it is used in everyday language, carries an emotional charge, and this emotional charge is the passion. Artificial intelligence is an attempt to realise the "meaning-in-itself" in the modern age, trying to ignore personal enthusiasm and passion. But passion-free invention cannot exist, and, for now this seems to remain a lasting if not eternal barrier to artificial intelligence.

[11] See Nietzsche (1968. 301): "A 'thing-in-itself' just as perverse as a 'sense-in-itself', a 'meaning-in-itself'. There are no 'facts-in-themselves', for a sense must always be projected into them before there can be 'facts'. The question 'what is that?' is an imposition of meaning from some other viewpoint. 'Essence', the 'essential nature', is something perspectival and already presupposes a multiplicity. At the bottom of it there always lies 'what is that for me?' (for us, for all that lives, etc.)".

IV. SCIENTIFIC INTEREST IS UNQUANTIFIABLE

Finally, let us examine the reason for the apparent contradiction that while AI systems can beat top-ranked minds in the mental sports and games mentioned in the introduction, they cannot produce groundbreaking novelties in the field of scientific discovery.

In chess, various calculation methods that assign numerical values to each piece and each step or position are well known (for example, according to classical piece value calculation, 1 unit is assigned to pawn, 3 to bishop, 5 to rook, 9 to queen). The purpose of the computer is to find a step that optimises the cumulative value of the current position. It can be done by examining an easy-to-construct mathematical tool, the so-called *objective function*, and to find its optimal value. A similar objective function – or in multi-criterion decision models, functions – can be defined in other games and in very different areas of the application of artificial intelligence as well, such as automated stock trading, where the obvious objective function is the amount of profit.

The objective function (or functions) must clearly quantify which of the two situations or states is more valuable, that is, when we make a better choice, to which direction belongs more *utility*. However, in the light of the above mentioned problems, it seems that such an objective function cannot be defined in scientific discovery.

It is worth mentioning here the notion of utility measured by a given objective function (a certain type of objective function is also called a utility function), because when we apply it to science, to scientific discovery, it brings to mind Kant's classical discussion of the conflict of faculties:[12]

> …*truth* (the essential and first condition of learning in general) is the main thing, whereas the *utility*… is of secondary importance (Kant 1992. 7).

Of course, this allows that what is useful may be untrue (and vice versa); meanwhile, usefulness and utility cannot be the primary guiding principle for a theoretical researcher. As Mihály Vajda expresses in his commentary on Kant's work above (Vajda 2016):

[12] This text is especially relevant to our topic because we may well assume that Kant, had it existed at that time, would have classified the Information Technology (or simply IT) faculty as one of the higher-utility faculties, in contrast to the lower faculties such as Philosophy (and Mathematics), where the guiding principles are pure erudition, free choice of subject, and critical approach.

> I would like to hope that the university will continue to train not only smart professionals but also a group of people who are carriers of something which is *per se* unreasonable, because it is useless… What can be contrasted with utility is a kind of irrationality: a world where the useless – beauty and tranquility – (also) reigns.

Scientists may choose a direction which is (at a given moment) seemingly useless, if they find this direction interesting. Moreover, accoding to Vajda's commentary, this – perhaps unreasonable – moment showcases the freedom and beauty of pure science.

Are we able to algorithmise and measure the motivation behind the drivers of what may seem like an unreasonable, useless choice? Recent scientific experiments show that we are powerless in this matter. Let us examine what happens when someone tries to define an "objective function of intrinsic interest", that is a metric stimulating and measuring the curiosity of an AI system based on the novelty of information or the amount of information obtained per unit of time.

In a recent experimental study (Burda 2018), authors found that the most interesting series of events (that is to say: the series providing the most interesting information defined by the objective function of intrinsic interest) to the AI system as it "watches" television is the continuous, instantaneous switching of channels, or even the black-and-white noise of the television (when there is no broadcasting), since the information per pixel changes the most in these cases. If, ironically, the computer was playing computer games to arouse its interest, it was observed that the computer – after several wins – sometimes "intentionally" lost the game in order to see "GAME OVER" which was rarely shown, so it was interesting new information according to the objective function. The irrationality here also appears, but the algorithmic uselessness is a dead end.

The above anomalies also prove that we cannot as yet properly allocate value to the interest of a process, situation, or even to the interest of an unknown scientific claim. While we can measure their information content in a technical and syntactic sense, we are unable to mathematise its semantic aspects, value, and intrinsic interest. Between stacked syllogisms, we are technically unable to set up a scale or order of values that can be automatically calculated and verified. All this eventually results in the AI system being able to function as a "smart professional" in the sense of Vajda, but – in the absence of a proper objective function – the system will be incapable of decision and choice, in the sense of Poincaré. Thus it remains entirely unclear if and how fully automatised scientific discovery could be carried out. Consequently, being a scientist full of passion and with the ability of selection remains humans' intellectual privilege.

REFERENCES

Alai, Mario 2004. AI, Scientific Discovery and Realism. *Minds and Machines*. 14. 21–42.
Bouzy, Bruno – Tristan Cazenave 2001. Computer Go: An AI Oriented Survey. *Artificial Intelligence*. 132. 39–103.
Burda, Yuri – Harri Edwards – Deepak Pathak – Amos Storkey – Trevor Darrell – Alexei Efros 2018. Large-Scale Study of Curiosity-Driven Learning. *arXiv:1808.04355*
https://arxiv.org/abs/1808.04355
Fehér M., István 2017. Prejudice as a Precondition for Understanding. *Hungarian Philosophical Review*. 61. 9–28 (in Hungarian).
Gadamer, Hans-Georg 2004. *Truth and Method*. Transl. by Joel Weinsheimer and Donald Marshall. London – New York/NY, Continuum Publishing.
Heidegger, Martin 1962. *Being and Time*. Transl. by John Macquarrie and Edward Robinson. London, SCM Press.
Kant, Immanuel 1992. *The Conflict of the Faculties*. Transl. by Mary J. Gregor. Lincoln, University of Nebraska Press.
Kauwe, Steven K. – Jake Graser – Ryan Murdock – Taylor D. Sparks 2020. Can Machine Learning Find Extraordinary Materials? *Computational Materials Science*. 174. 1–7.
King, Ross D. – Vlad S. Costa – Chris Mellingwood – Larisa N. Soldatova 2018. Automating Sciences: Philosophical and Social Dimensions. *IEEE Technology and Society Magazine*. 37. 40–46.
Kitano, Hiroaki 2016. Artificial Intelligence to Win the Nobel Prize and Beyond: Creating the Engine for Scientific Discovery. *AI Magazine*. 16. 39–49.
Lakatos, Imre 1976. *Proofs and Refutations*. Cambridge, Cambridge University Press.
Langley, Pat – Herbert A. Simon – Gary L. Bradshaw – Jan Zytkow 1987. *Scientific Discovery: Computational Explorations of the Creative Processes*. Cambridge/MA, The MIT Press.
Moravčík, Matej – Martin Schmid – Neil Burch –Viliam Lisý – Dustin Morrill – Nolan Bard – Trevor Davis – Kevin Waugh – Michael Johanson – Michael Bowling 2017. Deepstack: Expert-Level Artificial Intelligence in Heads-Up No-Limit Poker. *Science*. 356. 508–513.
Nietzsche, Friedrich 1968. *The Will of Power*. Transl. by Walter Kaufmann and John R. Hollingdale. New York, Vintage Books.
Poincaré, Henri 1912. *Science et Méthode*. Paris, Flammarion.
Poincaré, Henri 2009. *Science and Method*. Transl. by Francis Maitland. New York, Cosimo.
Polanyi, Michael 1961. Knowing and Being. *Mind*. 70. 458–470.
Polanyi, Michael 1974. *Personal Knowledge*. London, Routledge.
Popper, Karl R. 1950. Indeterminism in Quantum Physics and in Classical Physics. Part II. *The British Journal for the Philosophy of Science*. 1. 173–195.
Rapaport, William J. 2018. What is a Computer? A Survey. *Minds and Machines*. 28. 385–426.
Tshitoyan, Vahe – John Dagdelen – Leigh Weston – Alexander Dunn – Ziqin Rong – Olga Kononova – Kristin Persson – Gerbrand Ceder – Anubhav Jain 2019. Unsupervised Word Embeddings Capture Latent Knowledge from Materials Science Literature. *Nature*. 571. 95–98.
Ulam, Stanislaw M. 1976. *Adventures of a Mathematician*. New York, Charles Scribner's Sons.
Vajda, Mihály 2016. The Conflict of the Faculties. *Magyar Tudomány*. 177. 410–416 (in Hungarian).
Weber, Max 1946. Science as a Vocation. In Max Weber: *Essays in Sociology*. Transl. and ed. by Hans H. Gerth and Wright C. Mills. New York, Oxford University Press. 129–156.
Wilczek, Frank 2016. *Fantastic Realities: 49 Mind Journeys and a Trip to Stockholm*. Singapore, World Scientific.

Zsolt Kapelner

Why not Rule by Algorithms?*

Abstract

The rise of artificial intelligence (AI) poses new and pressing challenges to society. One such challenge is the increasing prevalence of AI systems in political decision-making, which is often considered as a threat to democracy. But what exactly is lost when certain aspects of political decision-making are handed over to AI systems? To answer this question, I discuss an extreme case in which all political decisions are made by intelligent algorithms that function without human supervision. I will call this case Rule by Algorithms. I consider the epistocratic argument for Rule by Algorithms according to which as long as algorithms can be expected to produce better outcomes than human rulers, we have a good reason to abandon democracy for algorithmic rule. Some authors attempt to resist such conclusions by appealing to the notion of public justification. I argue that these attempts to refute Rule by Algorithms are ultimately unsuccessful. I offer an alternative argument according to which Rule by Algorithms should be rejected because it imposes impermissible constraints on our freedom. The discussion of this extreme case provides valuable insights into the challenges of AI in politics.

Keywords: artificial intelligence, democracy, epistocracy, public reason, domination, freedom

* An earlier version of this paper was given at the 2019 workshop *Artificial Intelligence: Philosophical Issues*, part of the *Action and Context* series organized by the Department of Sociology and Communication, Budapest University of Technology and Economics (BME) and the Budapest Workshop for Language in Action, Department of Logic, Institute of Philosophy, Eötvös University (ELTE). This research was supported by the Higher Education Institutional Excellence Grant of the Ministry of Human Capacities entitled *Autonomous Vehicles, Automation, Normativity: Logical and Ethical Issues* at the Institute of Philosophy, ELTE Faculty of Humanities.

I. INTRODUCTION

Artificial intelligence (AI) systems exert a profound impact on today's society.[1] They fundamentally transform commerce, travel and communication, culture and learning. In recent years AI also started to shape government and public policy. Although one may argue that AI technologies contribute to the efficiency of political decision-making, many view the increasing prevalence of AI and automation in political decision-making as a threat to democracy. In this paper I discuss this problem by considering an extreme case. Imagine that at some point in the future intelligent algorithms that are able to function without human supervision completely take charge of political decision-making. What exactly would be objectionable about such an arrangement? What values would *Rule by Algorithms* undermine? To answer this question, I present in the first section the epistocratic argument for Rule by Algorithms, according to which if decision-making algorithms can be expected to make far better decisions than humans, then we have a good reason to replace democracy with Rule by Algorithms. Some authors might try to resist such a conclusion by appealing to arguments from *public reason liberalism*. In the second section I discuss this type of argument and conclude that it is ultimately unsuccessful. In the third section I present an alternative consideration against Rule by Algorithms, based on the concepts of freedom and domination.

II. THE EPISTOCRATIC CASE FOR RULE BY ALGORITHMS

It is widely accepted that political decisions ought to be made *democratically*, i.e. either directly voted on by citizens or authorized through some such vote.[2] Yet many argue that democracy should not enjoy this default status, and other forms of government might be preferable to it. One such proposed alternative is epistocracy, i.e. rule by experts or knowers. Its advocates argue that those who have expertise that pertains to political decision-making are in a better position to make *good* decisions than those who lack such expertise. Therefore, if political decision-making were left only to experts, its quality could be expected to increase. And, as Jason Brennan, one of today's leading advocates of epistocracy, argues, citizens have a fundamental right to competent government (Brennan 2011). Therefore, insofar as epistocracy can be expected to be more competent than democracy, we have a pro tanto obligation to replace the latter with the former.[3]

[1] For a discussion on the definition of, current research on AI as well as its potential future impact on society see Russell and Norvig (2016) and Boden (2016).

[2] For a discussion on the definition of democracy, see Waldron (2012) and Goldman (2015).

[3] For a more detailed discussion on epistocracy see Gunn (2014), Brennan (2016a), and Moraro (2018).

This epistocratic argument, if successful, provides a prima facie case for introducing a form of government in which political decisions are made exclusively by extremely intelligent and competent AI systems. If some AI systems could be developed which reliably emulated those cognitive and deliberative faculties by means of which humans make political decisions – except in a much faster, more accurate, and more efficient manner – and if these AI systems thereby attained a level of competence unavailable to humans or groups of humans, whether a democratic public or an expert panel, simply because of their natural limitations, then the epistocratic argument – being focused exclusively on the quality of political outcomes – would favour these AI systems as rulers over any human, expert or not.

One can conceive of Rule by AI in many different ways. For example, one may imagine that AI experts at one point create a superintelligent machine which then establishes itself as the robotic overlord of society (Bostrom 2014. 95). While these scenarios of superintelligent artificial dictators may be interesting, here I would rather focus on a different, no less fanciful, but perhaps somewhat more relevant case. AI systems, particularly machine learning algorithms are already in use in many areas of government and public policy. Such algorithms already support policymaking through data mining, they help optimizing the provision of public services, they provide risk-assessment data for criminal sentencing, they control traffic lights, and carry out many more tasks previously done by humans (Wirtz–Weyerer–Geyer 2019; Oswald 2018; Lepri et al. 2017; Coglianese–Lehr 2017). Suppose that as these algorithms become more sophisticated and efficient, we gradually hand over more and more tasks to them until all aspects of legislation, government, and perhaps even judicial tasks are handled by intelligent algorithms without human supervision. I call this case Rule by Algorithms.

I discuss Rule by Algorithms not because I believe that it can become reality anytime soon. My goal, rather, is to gain insight into the way in which the fundamental values of democracy can come into conflict with the increasing prevalence of AI systems in society and politics, and today the relevant type of AI system is closer to a machine learning algorithm than to a superintelligent digital dictator. Furthermore, certain core features of Rule by Algorithms are particularly interesting in comparison with the digital dictator scenario, as it will become clear in later sections.

There are a few assumptions I make about Rule by Algorithms here for the sake of the argument. First, I assume that Rule by Algorithms can be expected to produce significantly better outcomes than human decision-makers; otherwise the question of its preferability to democracy would not even arise. Second, I assume that ruling algorithms do not form a coherent mind or an *artificial person*[4]

[4] Here the term "artificial person" does not refer to the legal concept under which corporations and the like also count as artificial persons, but rather to a human-made AI system with

with its own interests, desires, volitions, beliefs, and so on. Rule by Algorithms, therefore, does not mean handing over power to a robotic overlord, but rather to a cluster of intelligent algorithms each carrying out various tasks pertaining to decision-making. The cooperation of these various algorithms emulates the way in which ordinary decision-makers produce outcomes, without constituting a coherent mind; in roughly the same way as various algorithms today (e.g., those used by social media sites or other online platforms) govern much of our lives without necessarily congealing into a single artificial patriarch overseeing our activities.[5]

Third, I assume that the algorithms would be sufficiently independent of their makers not to be thought of as mere tools in the hands of those who create them. Clearly, some human involvement is necessary for setting up and running Rule by Algorithms; someone has to make them, maintain them, etc. But for the scenario to be even worthy of discussion, the algorithms must be conceived of as being able to function on their own to a great extent, without human supervision. Their makers and users cannot have control over or ability to predict the outcome of the functioning of the algorithms in a precise manner.[6] This assumption is crucially important to distinguish Rule by Algorithms from AI-enhanced epistocracy, or from Rule by Software Engineers.

Assuming, then, that such algorithms could take over political decision-making, should we let them? One may argue that Rule by Algorithms is impossible. It requires human-level AI or Artificial General Intelligence (AGI), which according to many authors cannot be constructed (Boden 2016. 153–155). However, even if AGI is impossible – which it may not be (Turner 2019. 6n19) – it is not immediately clear that Rule by Algorithms requires AGI. A further argument is needed to show that to emulate all the deliberative faculties we use in political decision-making requires the artificial reproduction of the human mind in its entirety. Such a claim cannot simply be presupposed. And, in fact, it seems that in many areas of political decision-making which call for solving coordination problems and allocating resources efficiently – non-AGI-type – algorithms could be expected to do as good if not a better job than humans.

It is true, however, that there is more to political decision-making than solving coordination problems. Government also involves setting long-term goals and settling hard questions of value. But algorithms, the objection goes, could not do this on their own; such goals and core values would have to be ultimate-

all the features that constitute personhood in ordinary humans, e.g., a mind, the capacity for rational deliberation etc.

[5] One may object that any such cluster of algorithms would be bound to constitute a coherent mind and ultimately an artificial person. I will assume without argument that this is not the case, acknowledging that if it were, my account would have to be adjusted accordingly.

[6] For more on such algorithms and the ethical issues concerning them see Mittelstadt et al. (2016).

ly supplied by humans. Note, however, that the same is true of human decision-makers. Humans do not conjure long-term goals and values out of thin air; we are socialized by other humans, our reflections on values and goals start with material we receive from parents, teachers, and society in general. Still, as long as we operate on this material in a sufficiently independent manner, we can be thought of as making our own decisions. Similarly, perhaps humans would supply initial material on which algorithms operate for setting goals and making value-judgements; but as long as they function sufficiently independently, emulating those deliberative faculties humans use for setting goals and settling questions of value – which, again, cannot be simply stipulated to be impossible – they may be thought of as ruling on their own.[7] Thus, while human involvement would not be absent from Rule by Algorithms, ruling algorithms would not be mere pawns of any human being any more than human decision-makers are mere pawns of their parents or teachers.

This short discussion shows that there are no obvious reasons to discard Rule by Algorithms as in principle impossible. There may very well be nonobvious reasons, supported by further arguments, as well as reasons to think of it as practically unfeasible. Indeed, if, due to contingent circumstances, we never arrive at a level of technological advancement where Rule by Algorithms would be possible, reasonably inexpensive, and safe to implement, then introducing it in real life will never be an issue. But this does not affect the main argument of this paper, which is not about future scenarios for the use of AI in government, but rather the philosophical question of what kind of challenge, if any, is posed by AI to democracy. Rule by Algorithms is simply a hypothetical scenario which I use to draw out conclusions about this question; it can fulfil this role without ever being feasible in real life.

A final objection to consider is that the epistocratic argument presented above misunderstands the nature of political decision-making. It is false, one might claim, to say that there are better and worse decisions in politics, for political decisions are about values rather than facts, about clashes and compromises between antagonistic interests, rather than puzzles in social engineering where a solution can always be singled out as unambiguously optimal. Even if algorithms could emulate reasoning about goals and values, there is no sense in which their decisions could be better than those of humans, and therefore the epistocratic case for Rule by Algorithms evaporates.

This objection, again, relies on certain non-obvious premises which need to be argued for before the strength of the objection can be assessed. For example, even if in the case of certain value-judgements there is no way to tell which is

[7] Recall, again, that no single algorithm needs to have the capability to do all this on its own. It is sufficient if the collective functioning of all the ruling algorithms emulates these deliberative faculties without congealing into a single artificial mind.

better, it seems rather implausible to say that *no* distinction between better and worse decisions can be made when it comes to society's final goals and basic values. There is a clear sense in which Nazi Germany's choice of basic values and final political goals were much worse than many alternatives. As authors such as David Estlund (1993), Susan Hurley (2000) and Hélène Landemore (2012) argued, any plausible conception of politics must accept at least some degree of *political cognitivism*, i.e., the view that some political decisions, e.g., ones that promote liberty and prosperity, are better than others, e.g., those that promote destitution and tyranny, as an epistemically accessible objective matter of fact. With these considerations in mind, what should we think of the epistocratic case for Rule by Algorithms?

III. PUBLIC REASON AGAINST RULE BY ALGORITHMS

Some defences of democracy against epistocracy are epistemic in nature. Proponents of *epistemic democracy* argue, for example, that democracy possesses epistemic merits that epistocracy would lack, and is therefore in a better epistemic position to identify good political outcomes.[8] But since I assumed that Rule by Algorithms would outperform human decision-makers, the democratic answer to this challenge needs to be non-epistemic.[9] A prominent line of non-epistemic arguments against epistocracy comes from the tradition of *public reason liberalism*. Public reason liberals, such as John Rawls (1993) and Gerald Gaus (1996; 2010), argue that the exercise of political power is legitimate only if it is justifiable to all reasonable points of view, i.e., justified on the basis of reasons that are accessible to all reasonable members of society.[10]

David Estlund puts forward one of the most well-known arguments for the claim that epistocracy cannot be jusitified to all reasonable, or, as he calls them, qualified points of view (Estlund 2008. 48). Epistocracy justifies the exercise of coercive political power by appealing to the better outcomes that experts are able to produce due to their epistemic superiority. But as Estlund's *demographic objection* holds, "it is not unreasonable or disqualified to suspect that there will be other biasing features of the educated group, features that we have not yet identified and may not be able to test empirically, but which do more epistemic harm than education does good" (Estlund 2008. 222). For example, the experts may all come from wealthy families or be members of an otherwise dominant so-

[8] For more on epistemic democracy see Estlund (2003), Landemore (2012), and Peter (2016).

[9] There are practical objections to standard epistocracy as well, which I cannot discuss in detail here (Viehoff 2016; Arneson 2009).

[10] For more discussion on public reason liberalism, the criteria of reasonableness, public justification, and other related concepts see Chambers (2010) and Gaus (2015).

cial group which can make them biased in favour of their own group and against others. These distorting factors may detract from their ability to create good outcomes for everyone regardless of their epistemic superiority.

Estlund's argument is not that experts would surely produce biased outcomes. His argument is that it is not unreasonable to suspect that they would. It is also not unreasonable to reject this suspicion. Reasonable people can disagree about whether or not experts would produce the best outcomes. But precisely because this kind of reasonable disagreement is possible, the rule of experts cannot be justified to all qualified points of view potentially subjected to this rule by appealing to the consideration that experts would produce the best outcomes. Some could reject this consideration on reasonable grounds, and thus subjecting the population to the authority of experts would not be publicly justified.

One may argue against Rule by Algorithms in a similar way. Algorithms, however well they compute, can exhibit bias (Barocas–Selbst 2016; Howard–Borenstein 2018); therefore, one may argue that it is not unreasonable to suspect that although algorithms could produce good outcomes, the features that enable them to do so may travel with epistemically countervailing features that hinder this capacity. John Danaher (2016) formulates a related worry. He points out that algorithmic decision-making is often *opaque*, i.e. algorithms' decision-making mechanisms are not always transparent even to their makers or other experts. However, legitimate authority has a *non-opacity requirement*: decisions must be made based on reasons and principles that all reasonable or qualified citizens can endorse; if these reasons and principles cannot be accessed by citizens, not even in principle, then the decisions have no authoritative force and are illegitimate. For it is then never unreasonable for citizens to suspect that opaque decision-making mechanisms appeal to principles and reasons which they could reasonably reject (Danaher 2016. 251–252).

Note, again, that the argument does not presuppose that ruling algorithms are bound to be biased or to appeal to unacceptable reasons in their opaque decision-making. The argument only claims that it is not unreasonable to suspect that they would. Again, there may be reasonable disagreement on these worries. For example, reasonable people may argue that there are satisfactory safeguards against algorithmic bias which ultimately may even prove to be more successful in eliminating unfairness in political decision-making than any kind of human intervention (Zarsky 2011. 312; Zarsky 2016. 126). The point is that reasonable disagreement is possible on this matter, which undermines the legitimacy of Rule by Algorithms, as it undermines the legitimacy of standard epistocracy.

Are these arguments successful? I have my doubts. It certainly seems plausible that "as democratic citizens have the right to scrutinise and hold [to] account the exercise of political power, so algorithmic constituents have the right to scrutinise and hold account the exercise of algorithmic power" (Binns 2018. 553). But political power wielded by humans and algorithmic power wielded

by nonpersons without human supervision are fundamentally different. First, it is unclear if ruling algorithms would make and enforce rules in the same sense human holders of political power do, i.e., by creating authoritative directives which we have to obey. Algorithms on social media do not issue directives as to which advertisements we must watch or which news we must read. Rather, they shape our digital environment in such a way that we cannot help but act in certain kinds of ways (Yeung 2017). How similar this form of governance is to the traditional exercise of coercive political power is far from clear. But let us set aside this issue for the sake of the argument.

A more important difference between ordinary political rule and Rule by Algorithms is that the former requires that someone or some group be placed in a position of authority. In that position they are granted rights to treat others in ways which are prima facie *impermissible*. Normally we are not allowed to issue commands and coerce our fellow human beings to obey them. This would involve treating them not as moral equals, but as inferior beings subject to our private will. Only when the exercise of coercive power *by some persons over others* is publicly justified, is this threat averted. In other words, the demand of public justification is based on the fact that one person wielding coercive power over another carries an extremely high risk of moral injury, i.e., that of treating others as non-equals.

The reason why many authors endorse democracy is precisely because insofar as it distributes political power equally it does not threaten but rather affirms individuals' standing as equals (Christiano 2008). In a well-functioning democracy no one has the power to subject the polity to their private will, for citizens have an equal say in shaping political outcomes; in effect, no one rules over anyone (Kolodny 2014. 227). In contrast, non-democratic arrangements, e.g., epistocracy, as Estlund notes, "introduce an extra element of rule of some by others, and that element is subject to the qualified acceptability requirement, whereas its absence is not" (Estlund 2008. 219). Note, however, that the rule of some by others is also absent under Rule by Algorithms. In Rule by Algorithms the ruling is done by nonpersons, and therefore no one is threatened with being subjected to anyone else's private will.[11] Algorithms have no private will. For this reason, Rule by Algorithms seems to be on a par with democracy at least insofar as it also does not introduce the "extra element of rule of some by others".

This clarifies why I focus on the special case of Rule by Algorithms rather than Rule by AI more generally. If Rule by AI meant simply creating an artificial person with a single mind and extraordinary decision-making capabilities to rule over others, similarly to Bostrom's dystopia, then the case would not be significantly different from ordinary epistocracy. If artificial dictators are subject

[11] Recall the distinction between Rule by Algorithms and Rule by Software Engineers from the previous section.

to the same moral requirements as human ones, then it seems that Estlund's anti-epistocratic argument, if successful, would reject this form of Rule by AI as well.[12] The same is not true of Rule by Algorithms, however. Under Rule by Algorithms citizens are not ruled by anyone, for the ruling algorithms – not persons themselves – are sufficiently independent and intelligent not to be thought of as mere extensions of any person who have been involved in their making. Citizens are instead ruled by impersonal mechanisms which can be expected to reliably produce good political outcomes. Why should we reject such a proposition?

Of course, the claim that neither democracy nor Rule by Algorithms involves unjustified power hierarchies does not imply that one can never object to democratic or algorithmic decisions. If either democracy or Rule by Algorithms produced overt injustices, their rule should be condemned and resisted. Political decision-making, whether automated or not, should always take place in an institutional environment where strong, e.g., constitutional, guarantees guard against the worst injustices, protect human rights, individual liberty, and so on. Democracy, epistocracy and Rule by Algorithms should always be subjected to such restraints. Here we do not discuss the legitimacy of the unrestrained absolute dictatorship of algorithms, experts or majorities, as these are non-starters for any theory of legitimate authority.

Note that even if the dangers of absolute algorithmic dictatorship are averted, reasonable worries remain that Rule by Algorithms would not produce the best outcomes. But these alone, absent potentially problematic asymmetric power relations, are insufficient to disqualify Rule by Algorithms the same way they disqualify epistocracy. Democracy can be reasonably suspected not to produce the best outcomes as well. Even if one is convinced of the wisdom of the crowds, one is not necessarily compelled to think that crowds are always the wisest. Thinking that non-democratic forms of decision-making would be epistemically superior to democracy is not an unreasonable view. But the reasonable suspicion of epistemic suboptimality only has illegitimating force when it is coupled with the introduction of asymmetric power relations. Estlund's argument, in the end, is that reasonable people can suspect that under epistocracy they would sacrifice their political equality for nothing; for epistocracy may fail to deliver the great outcomes it promises. But what exactly would be sacrificed under Rule by Algorithms? Certainly not equality, which seems to trigger the demand of public justification in Estlund's argument. Then what? In the next section I will provide an answer to this question.

[12] Many authors challenge Estlund's argument (e.g., Lippert-Rasmussen 2012). Even if it fails, however, my point holds for the general approach of using public reason liberalism against Rule by Algorithms.

IV. DOMINATION BY ALGORITHMS

In my view, the test Rule by Algorithms fails is not that of equality but that of freedom. Political philosophers generally agree that political institutions should cater to citizens' freedom to some extent. For many authors, this means primarily the protection of certain basic liberties which carve out, for each citizen, spheres of non-interference within which said citizen's freedom can be exercised.[13] Ruling algorithms should, in principle, have no problem with securing these basic liberties for citizens and thus one might believe that there is no reason to think of them as particularly grave threats to freedom.

However, as philosophers – especially within the so-called *republican* tradition (Lovett–Pettit 2009) – pointed out, freedom is threatened not only by *interference* in what is usually thought of as individuals' *private affairs*, guarded by their liberty rights, but also by relations of *domination*. Domination is a complex idea, but it may be initially defined – although this definition will be revised later – as subordination to an alien will (Pettit 2012a. 79) exemplified most clearly by the relationship of the master and the slave. A slave remains subordinated to the will of the master even when said master decides out of benevolence never to interfere with the life of the slave.

Similarly, citizens may remain unfree in significant ways if certain subordination relations obtain, which may be the case even if their liberty rights are never breached. For example, even if a benevolent dictator – or a panel of experts under epistocracy – were to define liberty rights exactly in the desirable ways, e.g., granting free speech, free association, occupational freedom and all other important liberties, citizens may still be thought of as not having their freedom sufficiently protected, insofar as their liberties depend entirely on the benevolence of the dictator or the goodwill of the experts. Non-domination, as republicans often argue, requires not only that laws protect citizens' liberty rights, but also that citizens exercise *control* over legislation and political decision-making in general through democratic institutions (Pettit 2012b).[14]

Rule by Algorithms may seem immune to this challenge. Domination is most often thought of as a matter involving two persons, one dominated and one dominating. But as I noted so emphatically in the previous sections, algorithms are not persons, and as such, one may argue, they cannot dominate. I would dispute this point. In my view, domination or something very similar to it is possible without dominating agents. Some authors forcefully deny this claim (Lovett 2010. 47–49), while others point to oppressive social structures, such as

[13] A view roughly along these lines is presented, for example, in Rawls's *A Theory of Justice* (Rawls 1999. 177–178).

[14] Note that in this paper I do not endorse the republican doctrine that domination is the *only* threat to freedom, and that there is nothing more to freedom than non-domination. I only claim that non-domination is an important aspect of freedom.

patriarchy or perhaps capitalism as potential non-agential sources of domination (Gourevitch 2013; Einspahr 2010). I would like to focus on a different case of non-agential domination proposed by Gwilym Blunt (2015), who invites us to imagine a scenario in which an unjust apartheid system is set up

> by a legislator who then promptly dies. The laws are impartially enforced, not by the privileged group, but by a series of automatons; they enforce the law impartially and cannot be reprogrammed. In this case, all groups have no influence over their status, even though one group is privileged, they cannot be said to dominate the others since they do not have systemic or interactional arbitrary power. They do not even act as agents of domination. The automatons cannot be said to dominate since they are not agents, but only machines with no will of their own. The legislator cannot be said to dominate after laying down the law, since he is dead and has no agency. It seems at least possible that this would be an instance of 'pure' systemic domination. (Blunt 2015. 18)

One may dispute that this is in fact domination. Still, it is clearly a worrying case of restricting individuals' freedom. Even if it were shown that some terms other than domination would be more appropriate for the analysis of this case, this would not affect my argument greatly, as it is not premised on a domination-only view of freedom. For this reason, I will continue using the term "domination", acknowledging that if other terms are proven to be more suitable for this discussion, they should be adopted instead.

Now imagine that in Blunt's example the automatons are in fact algorithms that do not uphold an unjust apartheid system, but rather produce good outcomes; furthermore, they do not enforce the will of a deceased legislator, but rather act upon their own determination as autonomous AI systems. Would domination still obtain? In other words, do those characteristics of Blunt's example that engender domination carry over to Rule by Algorithms as I describe it? In my view, they do. The automatons in Blunt's example do not dominate because they uphold an unjust apartheid system. As the example of the benevolent dictator shows, domination may obtain under a relatively just system as well; the dictator may defend basic liberties, introduce fair distributive policies, and so on – still, citizens remain dominated under the dictator's rule.

The automatons also do not dominate because they carry out the will of a deceased person. Although previously I defined domination as subordination to an alien will, it is important to see exactly what it is about such subordination that threatens freedom. Michael Blake notes in a seminal paper that domination violates "the autonomy of the individual by replacing that individual's chosen plans and pursuits with those of another" (Blake 2001. 272). This replacement or substitution of wills seems to be an essential element of domination. There are two aspects to domination so understood, however. One is the *muting* of the individual's will, the rendering ineffective of the dominated person's choice to

pursue certain plans. The other is the replacement of that will with the will of someone else. I would argue that the first is sufficient for domination, or at least for the type of loss of freedom that is relevant both in Blunt's example and in the case of Rule by Algorithms.

Imagine an evil scientist who implants a chip in my brain. Every time I would make a decision about my career, for example, the chip turns off my deliberative faculties and selects a choice randomly. The scientist does not choose my career for me; it is possible that the scientist and I never cross paths again, she has no idea how I live my life and has no way of interfering with my choices anymore. She does not control me in any sense of the word, nor does she replace my will with hers; rather she merely mutes my will handing over this aspect of my life to pure chance. Still, this clearly subtracts from my freedom. I cannot pursue occupations that I find rewarding, establish work-life balance on my own terms, and so on, for my will regarding these matters is rendered weightless.[15]

In some cases, therefore, we can be unfree simply because a situation is so arranged that our will does not matter. This doesn't require that anyone else's matter *more*. There may be no one whose will does, and yet we remain unfree. Could intelligent algorithms render us dominated or at least unfree in this sense? They certainly could. Under Rule by Algorithms the rules that govern the shared life of the polity are made without the contribution of citizens; indeed, they are made without the contribution of anyone. Citizens' will regarding the terms of social cooperation is entirely irrelevant, impersonal mechanisms take care of settling all political matters.

Naturally, one is not always unfree when one's will does not matter. It does not matter that I wish to be able to breathe in outer space without aid; I simply cannot. This fact does not detract from my freedom one bit. As Isaiah Berlin famously remarks, "mere incapacity to attain a goal is not lack of political freedom" (Berlin 2002. 169). But citizens' inability to influence political decisions under Rule by Algorithms is not mere incapacity. They are subject to an artificial, human-made arrangement that they have brought into being and are able to change, even if that takes a major effort. Under this arrangement, although there is no one ruling over them, citizens also have no opportunity to weigh in on certain decisions which, many argue, would be crucially important for establishing their standing not only as equal, but also as free citizens of society. For as Ronald Dworkin notes,

[15] None of this is to say that the scientist should not be held *responsible* for what has happened to me. It seems very plausible to me that the scientist is indeed responsible and should be blamed for all the misfortunes that ensued from her operation. It is the scientist's *fault* that I am unfree, but I am not unfree because she substitutes her will for mine.

> We cannot make our political life a satisfactory extension of our moral life unless we are guaranteed freedom to express our opinions in a manner that, for us, satisfies moral integrity. […] But the demands of agency go beyond expression and commitment. We do not engage in politics as moral agents unless we sense that what we do can make a difference, and an adequate political process must strive, against formidable obstacles, to preserve that potential power for everyone. (Dworkin 2002. 201–202)

Delegating all political decisions to impersonal mechanisms would threaten our standing as free moral agents capable of and entitled to reason and make choices about the most fundamental aspects of our shared life in society. This, I believe, is sufficient reason to resist transition to Rule by Algorithms.[16]

If this is right, then the reason why Rule by Algorithms should be rejected is not that we can raise reasonable doubts about its effectiveness but that it imposes impermissible constraints on our freedom. For freedom is not something we can trade for greater economic efficiency or growth, more innovation, or whatever else is promised by the superior political decision-making of Rule by Algorithms. Rule by Algorithms threatens to take away control over certain aspects of our shared social and political life without which we cannot view ourselves as free and equal citizens living in a just society.

V. CONCLUSION

I do not believe that algorithms will rule us anytime soon. Still, there are important conclusions to be drawn from the discussion above. The rise of algorithmic decision-making, not only in government and politics, but also in other areas of life, raises innumerable questions and problems. Some of these, such as opacity or bias, serious as they are, can be expected to be mitigated via technical and institutional solutions (Lepri et al. 2018; Danaher 2016. 258–265). Others, however, force us to carefully reflect upon the basic principles and values according to which we organize society. The automation of decision-making through AI systems, including intelligent algorithms, promises to increase efficiency and the quality of outcomes if we are willing to give up control. Control over certain aspects of our lives, however, is constitutive of our freedom both as private individuals and as democratic citizens.

[16] Some authors doubt that democratic participation rights are in fact constitutive of citizens' freedom in society (Brennan 2016b). For a defence of the position that they are, see Gould (1990), Hanisch (2013), and Rostbøll (2015).

We need to understand what we risk when we hand over some or all of these aspects to impersonal mechanisms which promise to take better care of us than we ourselves could. Note that this promise does not have to be false. Decision-making algorithms may prove to be good stewards to our interests, even outstanding ones. And yet we may incur serious losses for which the gifts of hypercompetent algorithmic government, full or partial, may not be able to compensate. Here I discussed one such potential loss, i.e., the loss of freedom, and suggested that the introduction of forms of algorithmic decision-making that threaten with this kind of loss should be resisted. This is an important insight even if full Rule by Algorithms is but a distant possibility, for even the partial automation of political decision-making can diminish democratic control in ways which threaten citizens' freedom. Further research is needed for spelling out how these considerations translate into more tangible regulatory measures. Still, I hope that I made some steps toward the right direction and drew attention to some of the important aspects of these problems.

REFERENCES

Arneson, Richard J. 2009. The Supposed Right to a Democratic Say. In Thomas Christiano – John Christman (eds.) *Contemporary Debates in Political Philosophy*. London, Blackwell. 197–212.

Barocas, Solon – Andrew Selbst 2016. Big Data's Disparate Impact. *California Law Review*. 104. 671–732.

Berlin, Isaiah 2002. Two Concepts of Liberty. In Henry Hardy (ed.) *Liberty*. Oxford, Oxford University Press. 166–217.

Binns, Reuben 2018. Algorithmic Accountability and Public Reason. *Philosophy and Technology* 31. 543–56.

Blake, Michael 2001. Distributive Justice, State Coercion, and Autonomy. *Philosophy and Public Affairs*. 30. 257–296.

Blunt, Gwilym David 2015. On the Source, Site and Modes of Domination. *Journal of Political Power*. 8. 5–20.

Boden, Margaret A. 2016. *AI: Its Nature and Future*. Oxford, Oxford University Press.

Bostrom, Nick 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford, Oxford University Press.

Brennan, Jason 2011. The Right to a Competent Electorate. *Philosophical Quarterly*. 61. 700–724.

Brennan, Jason 2016a. *Against Democracy*. Princeton/NJ, Princeton University Press.

Brennan, Jason 2016b. Democracy and Freedom. In David Schmidtz – Carmen E. Pavel (ed.) *The Oxford Handbook of Freedom*. Oxford, Oxford University Press. 335–349.

Chambers, Simone 2010. Theories of Political Justification. *Philosophy Compass*. 5. 893–903.

Christiano, Thomas 2008. *The Constitution of Equality: Democrtic Authority and Its Limits*. Oxford, Oxford University Press.

Coglianese, Cary – David Lehr 2017. Regulating by Robot: Administrative Decision Making in the Machine-Learning Era. *Georgetown Law Journal*. 105. 1147–1223.

Danaher, John 2016. The Threat of Algocracy: Reality, Resistance and Accommodation. *Philosophy and Technology*. 29. 245–268.

Dworkin, Ronald 2002. *Sovereign Virtue*. Cambridge/MA, Harvard University Press.
Einspahr, Jennifer 2010. Structural Domination and Structural Freedom: A Feminist Perspective. *Feminist Review*. 94. 1–19.
Estlund, David 1993. Making Truth Safe for Democracy. In David Copp – Jean Hampton – John Roemer (eds.) *The Idea of Democracy*. Cambridge, Cambridge University Press. 71–100.
Estlund, David 2003. Beyond Fairness and Deliberation: The Epistemic Dimension of Democratic Authority. In Thomas Christiano (ed.) *Philosophy and Democracy: An Anthology*. Oxford, Oxford University Press. 69–94.
Estlund, David 2008. *Democratic Authority*. Princeton, Princeton University Press.
Gaus, Gerald 1996. *Justificatory Liberalism*. Oxford, Oxford University Press.
Gaus, Gerald 2010. *The Order of Public Reason*. Cambridge, Cambridge University Press.
Gaus, Gerald 2015. Public Reason Liberalism. In Steven Wall (ed.) *The Cambridge Companion to Liberalism*. Cambridge, Cambridge University Press. 112–140.
Goldman, Alvin 2015. What Is Democracy (and What Is Its Raison D'Etre)? *Journal of the American Philosophical Association*. 1. 233–256.
Gould, Carol 1990. *Rethinking Democracy: Freedom and Social Cooperation in Politics, Economy, and Society*. Cambridge, Cambridge University Press.
Gourevitch, Alex 2013. Labor Republicanism and the Transformation of Work. *Political Theory*. 41. 591–617.
Gunn, Paul 2014. Democracy and Epistocracy. *Critical Review*. 26. 59–79.
Hanisch, Christoph 2013. An Autonomy-Centered Defense of Democracy. *International Philosophical Quarterly*. 53. 371–384.
Howard, Ayanna – Jason Borenstein 2018. The Ugly Truth About Ourselves and Our Robot Creations: The Problem of Bias and Social Inequity. *Science and Engineering Ethics*. 24. 1521–1536.
Hurley, Susan 2000. Cognitivism in Political Philosophy. In Roger Crisp – Brad Hooker (eds.) *Well-Being and Morality: Essays in Honour of James Griffin*. Oxford, Oxford University Press. 177–208.
Kolodny, Niko 2014. Rule Over None I: What Justifies Democracy ? *Philosophy and Public Affairs*. 42. 195–229.
Landemore, Hélène. 2012. *Democratic Reason*. Princeton, Princeton University Press.
Lepri, Bruno – Nuria Oliver – Emmanuel Letouzé – Alex Pentland – Patrick Vinck 2018. Fair, Transparent, and Accountable Algorithmic Decision-Making Processes: The Premise, the Proposed Solutions, and the Open Challenges. *Philosophy and Technology*. 31. 611–627.
Lepri, Bruno – David Sangokoya – Emmanuel Letouzé – Nuria Oliver 2017. The Tyranny of Data? The Bright and Dark Sides of Data-Driven Decision-Making for Social Good. In Tania Cerquitelli – Daniele Quercia – Frank Pasquale (eds.) *Transparent Data Mining for Big and Small Data*. Berlin, Springer. 3–24.
Lippert-Rasmussen, Kasper 2012. Estlund on Epistocracy: A Critique. *Res Publica*. 18. 241–258.
Lovett, Frank 2010. *A General Theory of Domination and Justice*. Oxford, Oxford University Press.
Lovett, Frank – Philip Pettit 2009. Neorepublicanism: A Normative and Institutional Research Program. *Annual Review of Political Science*. 12. 11–29.
Mittelstadt, Brent Daniel – Patrick Allo – Mariarosaria Taddeo – Sandra Wachter – Luciano Floridi 2016. The Ethics of Algorithms: Mapping the Debate. *Big Data & Society*. 3. 1–21.
Moraro, Piero 2018. Against Epistocracy. *Social Theory and Practice*. 44. 199–216.

Oswald, Marion 2018. Algorithm-Assisted Decision-Making in the Public Sector: Framing the Issues Using Administrative Law Rules Governing Discretionary Power. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. 376. 1–20.

Peter, Fabienne 2016. The Epistemic Circumstances of Democracy. In Michael S. Brady – Miranda Fricker (eds.) *The Epistemic Life of Groups: Essays in the Epistemology of Collectives*. Oxford, Oxford University Press. 133–49.

Pettit, Philip 2012a. Freedom. In David Estlund (ed.) *The Oxford Handbook of Political Philosophy*. Oxford: Oxford University Press. 77–97.

Pettit, Philip 2012b. *On the People's Terms: A Republican Theory and Model of Democracy*. Cambridge, Cambridge University Press.

Rawls, John 1993. *Political Liberalism*. New York, Columbia University Press.

Rawls, John 1999. *A Theory of Justice*. Cambridge/MA, Harvard University Press.

Rostbøll, Christian 2015. The Non-instrumental Value of Democracy: The Freedom Argument. *Constellations*. 22. 267–278.

Russell, Stuart – Peter Norvig 2016. *Artificial Intelligence: A Modern Approach*. 3rd ed. New York, Pearson Education.

Turner, Jacob 2019. *Robot Rules: Regulating Artificial Intelligence*. London, Palgrave Macmillan.

Viehoff, Daniel 2016. Authority and Expertise. *Journal of Political Philosophy*. 24. 406–426.

Waldron, Jeremy 2012. Democracy. In David Estlund (ed.) *The Oxford Handbook of Political Philosophy*. Oxford, Oxford University Press. 189–203.

Wirtz, Bernd – Jan C. Weyerer – Carolin Geyer 2019. Artificial Intelligence and the Public Sector – Applications and Challenges. *International Journal of Public Administration*. 42. 596–615.

Yeung, Karen 2017. 'Hypernudge': Big Data as a Mode of Regulation by Design. *Information Communication and Society*. 20. 118–136.

Zarsky, Tal 2011. Governmental Data-Mining and Its Alternatives. *Penn State Law Review*. 116. 286–330.

Zarsky, Tal 2016. The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making. *Science Technology and Human Values*. 41. 118–132.

Contributors

Zsuzsanna Balogh, PhD. Philosopher, assistant professor at the Institute of Philosophy, Faculty of Humanities, Eötvös University (ELTE). She earned her PhD at CEU and her MA and BA degrees at the University of London. Within the *Autonomous Vehicles, Automation, Normativity: Logical and Ethical Issues* research project at ELTE Philosophy, she focuses on ethical and philosophy of mind-related aspects of AI systems.

Tomislav Bracanović is a senior research associate at the Institute of Philosophy in Zagreb, Croatia. He published two monographs – *Evolutionary Theory and the Nature of Morality* (2007) and *Normative Ethics* (2018), both in Croatian – and a number of book chapters and articles in scholarly journals. His research interests lie in ethics, applied ethics, philosophy of biology and philosophy of science and technology.

Miklós Hoffmann is full professor of Mathematics and Computer Science at Eszterházy Károly University and University of Debrecen. He received his PhD from University of Debrecen in 1998, and earned his DSc degree in 2016. Besides his work in mathematics and artificial neural networks, his current research focuses on various educational, ethical and philosophical aspects of these fields.

Zsolt Kapelner is currently pursuing a PhD in philosophy at Central European University. He is junior research fellow at Institute of Philosophy, ELTE, working within the research project *Autonomous Vehicles, Automation, Normativity: Logical and Ethical Issues*. His research areas include moral and political philosophy, in particular democratic theory, as well as global justice and critical theory.

Fabio Tollon has an MA in Philosophy from Stellenbosch University. His research interests are philosophy of technology, ethics of AI, and the epistemic status of delusional beliefs.