Knowing your robot: the fiduciary program in the age of AI

Mihály Héder

To a longtime supporter of our department in Budapest,
to a friend,
to Phil Mullins.

Introduction

Michael Polanyi is one of the most influential philosophers of science of the twentieth century, although perhaps he is not as well known as he is important. While his older brother, Karl, made a highly impactful, well-known contribution to the understanding of macroeconomics in his seminal work *The Great Transformation* ([1944] 1957), Michael Polanyi made a similarly impactful contribution to, well, all areas of thinking, really, with *Personal Knowledge: Towards a Post-Critical Philosophy* (1958).

This contribution, besides being nutritious food for thought for philosophers, also influenced the everyday work of engineers, designers, and managers, even if only indirectly. Sometimes we can even find cases of direct influence, as in the case of Xerox Parc.[1] The concept of *tacit knowledge* appears in the everyday considerations of engineering and bureaucracy and is cited almost as often today as in any of the other five decades that have passed since publication.

One wonders if the continued relevance of post-critical philosophy, which is apparent in the citations, is because Michael Polanyi's central message is still unheeded; the perception of a person's epistemic faculties and ontological status is still as distorted, at least in the grand scheme of things, as it was in the late 1950s, to which Polanyi responded. I will argue that this is indeed the case: the same self-defeating variant of critical approach Polanyi tries to surpass is still prevalent in the contemporary debates. One such debate is about the epistemic faculties and the moral and ontological status of robots, that is, the *machine question* (Gunkel 2012).

As to the source of this framing problem, the clue is in Michael Polanyi's *Personal Knowledge* subtitle: the overly zealous *critical* philosophy.

The entire post-critical project is often seen by many as a major contribution mostly to epistemology, but I argue that, while this is true, epistemology is a means to an end; it is applied to issues about organization of society and, ultimately, to profound moral questions, which are all part of Polanyi's message.

Post-critical philosophy

There is no better source for understanding the evolution of Polanyi's thinking on the matter than Phil Mullins (2001). He explains the origin of the

term “post-critical philosophy” and, more widely, the vision Polanyi offers. Mullins traces the term back to 1951, to the first series of the so-called Gifford Lectures. At the center of it is an emphasis on the indispensable personal participation of the knower in the known.

I will argue in this paper that research on our personal involvement in evaluating artificial intelligence (AI) will prove highly fruitful and insightful. Mullins positions Polanyi as a thinker who was trying to process the implications of the findings of Gestalt psychology, in that the whole perception cannot be reduced to the sum of its parts. I intend to frame the evaluation of AI with this in mind.

We also learn that by “critical” Polanyi means “Descartes, Hume, Kant, J. S. Mill, and Bertrand Russell” and the great success their systematic doubt brought to science. However, he also believes that this approach is reaching its limits as a heuristic, and a new perception is required about the methodology of scientific thinking. This is post-critical philosophy, which, in terms of history of science, can be called postmodern thought as well. However, as Mullins cites several sources on the matter, postmodern in this case is quite constructive and commitment-oriented, contrary to the usual associations of the term. But post-critical thinking tries to supersede not only unbounded doubt as the central heuristic of science but also the ideal of impersonal knowledge. To do that, Mullins explains, Polanyi rehabilitates confidence in “overt belief” (Mullins 2001, 78).

Together with the active involvement of the person, and the role of commitment, post-critical thinking also opposes all kinds of dualism without itself becoming monistic; through the theory of emergence, it rejects the dichotomy, and therefore the dilemma, between monistic and dualistic worldviews, and Mullins guides us to the works of Sanders (1991) and Gill (2000) on the matter.

The stakes

Before we can arrive at the question of AI, we need to understand just how high Polanyi’s aspirations were when he wrote *Personal Knowledge*. Starting from understanding an epistemic-ontological issue, he reframes the history of science, not only for deepening our knowledge of past events but also to reestablish the methodology of science. In his historical approach he even precedes Thomas Kuhn. At the same time, he derives some prescriptions, not only for science narrowly but also for social organization. He does all this not out of grandeur but because the acknowledgment of the tacit dimension starts a domino effect and logically demands the reframing of these issues—and Polanyi does not shy away from this task. Again, Mullins (2021) is our best guide on this topic.

The reason why this is important to us is that the multidimensional problem of the human reception of the sociotechnical system we call AI also necessitates an approach that is no less holistic or fiduciary, and the stakes of our approach to AI knowledge have a similar structure as those of our approach to scientific knowledge.

Before arriving at the moral issues related to AI, let us review some key points Polanyi has to say about morality in general. He was contemporary to several dark moments. Born in 1891 in Budapest, he served in the First World War and witnessed the revolutions and counterrevolutions that came afterwards, forcing him to emigrate to Karlsruhe, Germany. He then had to emigrate again, this time running from the Nazi party. The Second World War found him in Manchester, as a professor in physical chemistry. His students, Eugene Wigner and Melvin Calvin, won Nobel prizes, as did his son, John C. Polanyi.

Polanyi was a medical doctor who turned to physical chemistry and became a world leader in that field. But that never stopped him from reflecting on all kinds of issues outside of that field—he was a founder of the Galileo Circle, for instance. In Weimar, Germany, he experienced hyper-inflation, which made him interested in economics. Although he was less known for his interest in economy than his brother, he still made his own contributions to the field (Bíró 2019).

More crucial to this article is his experience of a trip to the Soviet Union. In 1935, he was invited to give lectures on scientific topics, but he also met with a main ideologist of Stalin, Bukharin, who explained to him his views on planned science versus the pursuit of pure science. Pure science, for Bukharin, was nothing more than a bourgeois pastime; he saw nothing of value in the freedom of scientists whom he believed should subsume their own work under the current five-year plan. Polanyi, on the other hand, had always been a defender of freedom of liberty in general and freedom of scientific pursuit in particular.

Bukharin's bloody execution, just a few years later in Stalin's next wave of purges, must have been a stark reminder for Polanyi of the importance of these values. Examples like this could have played a major role in his efforts to understand how smart, educated, successful scientists and individuals could support totalitarian regimes. Bukharin and others were by no means ill-informed people who were susceptible to being brainwashed by the cheap tricks of constant propaganda. So Polanyi became interested in how totalitarian regimes could achieve such a powerful grip and what science's part was in all this.

Moral inversion

Let's go through Polanyi's main trail of thought the unusual way, starting from the explanandum! What Polanyi witnessed was uncontrolled, unchecked revolution and sometimes raging destruction that was conducted with heated passion. It was the passion that troubled Polanyi the most: it appeared to be deeply moral in nature, yet the people experiencing it denied the existence of any kind of moral values.

To give the phenomenon a name, he came up with *moral inversion*. Moral inversion means that someone arrives at an intellectual position that replaces the pursuit of moral values—that search deemed irrational if one a priori postulates that moral values do not exist in the first place—with the pursuit of the objective truth. However, the moral passion that is still a necessary part of the human psyche—more on that later—does not actually disappear; it just becomes unchecked, unacknowledged, and much more dangerous.

This is most vividly evident, according to Polanyi, in the dynamics of communist movements. Nevertheless, Polanyi also references the insights of Friedrich Meinecke, who examined the intricate web of German power politics leading up to the First World War. Meinecke's analysis suggests that the Germans were initially conscious of the inherent moral hazards of wielding power. This awareness, however, did not prevent them from equating power with law. In doing so, they contended that any deviation from this alignment was merely an indicator of a misguided morality (Paksi and Héder 2020).

To put it more plainly, the Germans refuted the importance of traditional moral standards. In their view, by rejecting these standards, they could put forth their actions as morally upright, citing unfiltered sincerity as their justification. This approach saw them challenge what they perceived as the deceptive moralities of the Anglo-Saxons, aiming to expose and dismantle them. The underlying goal was to establish a moral high ground.

A similar contradictory dynamic is observable in Marxist ideologies and actions. Marxists, too, attempt to claim a moral superiority. They do this by disavowing traditional morality, but in their case they offer a comprehensive ideological rationale for their actions.

All of the above is of course very prima facie contradictory without an additional element. The way these ideologies work is that they claim to be more honest about the purported non-existence of these values and by this honesty become more virtuous. Of course, acknowledging that honesty about the non-existence of values is a virtue in itself would be another contradiction, but this circularity may remain hidden under rhetoric and obfuscation. This is an example of what Polanyi calls a "*deceptive substitution*" (Paksi and Héder 2020). The structure of such a move is quite straightforward. It means that a person thinks that certain beliefs (i.e., a

tacit mathematical intuition based on a sense of aesthetics) or values are not compatible with another, more important part of their belief system (i.e., objectivism) and therefore the former needs to be discarded. However, as Polanyi explains at length, beliefs or values cannot simply be discarded; hence, they are rebranded by the person as something else. This serves as self-deception primarily and appears deceptive to others as a side effect.

For instance, according to Polanyi, the convincing power of Marxism rises from its ability to resolve this tense contradiction: "[Marxism] enables the modern mind, tortured by moral self-doubt, to indulge its moral passions in terms which also satisfy its passion for ruthless objectivity" (Polanyi 1958, 228). He emphasizes that humans are not mere mechanical entities but beings pulsating with emotions and sentiments (this needs some justification, which comes later). Interestingly, even those who dismiss the idea of moral passions, advocating for a purely scientific objectivism, are ironically propelled by the very moral passions they deny.

Moral inversion can never be complete, however; therefore, it appears in gradual steps. *Disguised moral inversion* means that on the semantic level the values are challenged but on the practical level they are still followed. "Men may go on talking the language of positivism, pragmatism, and naturalism for many years, yet continue to respect the principles of truth and morality which their vocabulary anxiously ignores" (Polanyi 1958, 233).

One of his examples is Freud, who attempted to naturalize the moral passions of humans, that is, to reduce them to psychological factors, but never tried to achieve any kind of political change based on this. Instead, Freud strived to be a perfect, honorable gentleman who wanted nothing more than prosperity to all. About him and similar scientists, Polanyi ironically remarks, "Indeed, a writer who has proved his hard-headed perspicacity by denying the existence of morality will always be listened to with especial respect when he does moralize in spite of this" (Polanyi 1958, 234).

Then, an example of the less disguised form of moral inversion is the Marxist interpretation of science and the followers of it, among them Bukharin (Hartl 2021). They attempt to unmask and abolish the so-called bourgeois science to install something more "useful" for socialism at the expense of the independence of science. This process, depending on how long it was allowed to run its course, destroyed the actual scientific potential of the Soviet political system.

The concept of moral inversion works only if moral realism is also the case. Polanyi firmly believes in the existence of the inherent moral passions of humans, but he wants to do more than just postulate the existence of them.

As we continue to reverse-engineer Polanyi's worldview, this is where we see the true significance of his better-known concepts. One is tacit knowledge; the other is emergence. The concept of tacit knowledge, famously, concisely captures the fact that we know more than we can tell; that is, we are unable to rely on exclusively explicit inferences when we do science or, indeed, when we perform in any role at all. The function of this concept—apart from its inherent importance and practical consequences—for Polanyi's larger effort is to moderate our expectations regarding what can be achieved by explicit reasoning, including reasoning about the laws of nature, mathematics, or the *values of a society*. Also, recognition of the ubiquitous nature of tacit knowledge makes any performance—anything we are able to achieve at all—an example of an act during which we *integrate* tacit, momentarily uninvestigated, unreflected knowledge with explicit knowledge. This inevitably requires a sort of blind trust in the former. If we were actually to investigate and reflect on all elements of our performance —be it riding a bicycle, swimming, playing the piano, or conducting scientific research to the highest standards—the performance would fall apart.

Polanyi's titles such as the "fiduciary program" and "post-critical philosophy" call for recognition of the always present, "tacit dimension" with similar firmness as with which one accepts "cogito ergo sum." In other words, Polanyi believes to have shown—and good counterarguments are yet to be found—the undeniable universality of the tacit in animal and human action, including knowing. (Polanyi, just as in his native Hungarian language, does not use different words for "can" and "know how to." Knowing, that is, discovering reality and truth, for him shares the same principles as does any other action.) It follows that everyone has to have a level of *faith* in their own tacit faculties if they are to perform successfully, whether each person acknowledges it or not. Of course, this insight has tremendous scientific, engineering, and management consequences, for which Polanyi is best known.

Again, this is not presented as just a hypothesis—Polanyi believed that he offered a sufficient amount of crucial evidence to support that personal knowledge works the way he describes. Of course, the fiduciary nature of rationality cannot be explicitly proven—in fact, nothing can be truly explicitly proven; that would be a contradiction—so it has to be accepted based on the nearly endless examples Polanyi offers. That is, accepting the fiduciary nature of rationality itself is a fiduciary act, a leap in knowledge. This, at least, is not logically contradictory. *This approach does not live up to the objectivity standards set by critical philosophy. But, again, nothing else does either, as this standard is impossible to achieve; this is what makes the*

post-critical approach necessary. Yet the value of the critical standard as a heuristic and a driver for the scientific revolution cannot be overstated.

To explain how knowledge is truly generated and transmitted (if it cannot be fully explicated) Polanyi brings in his experience in radiology class at the Medical University. He recounts that the professor, despite his best efforts, was unable to give an explicit description of what the students needed to look for in lung X-rays; instead, he talked in seemingly vague concepts. However, the pupils specialized in radiology learned what to see and what to call things in the ostensive setting and to prove themselves by accurately identifying the relevant features on images that contained confirmed information, previously unknown to them, thereby becoming accepted as experts in the field of radiology. Such is the way we have to understand the nature of science or the values of societies.

Of course, the fiduciary act is a vastly fallible way of conducting anything, but there is no alternative. This fallibility is unavoidable and, in the grand scheme of things, moderated only by evolution (another huge topic of personal knowledge), which ensures that only those survive who can successfully and often enough assess reality.

The other pillar of Polanyi's arc of argument is emergence. A good entry point to this concept is the nature of the person that wields tacit knowledge. Polanyi argues that the person is not reducible to the concepts known to physics or to other branches of science. These descriptions of the world do not contain any teleological building blocks needed for that reduction. This is not only true for humans, either—Polanyi postulates the existence of operational principles that work on a higher level than physical and chemical laws. These govern machine-like entities like machines (as in the everyday use of the word) and living beings. The operational principles are necessary to meaningfully explain the behavior of these, but the lower-level conditions need to be within certain parameters for them to come into effect (Héder 2019). Just as with the tacit dimension, in this ontological matter Polanyi also proposes arguments that he believes to be ultimately convincing, albeit these have been the subject of much more criticism by posterity than the idea of the tacit (Paksi 2017). Yet the acceptance of some form—even if it is just a weaker one (Bedau 1997)—is quite common. Polanyi proposes a strong, ontological emergence, to support which his arguments touch on the irreducibility of the self (mind included) and the futility of the reduction of the teleological nature of certain entities.

The connection of all this to moral inversion is the following: what Polanyi essentially says is that while it is true that we cannot explicate our knowledge about moral values, it does not follow that they do not exist. The critical-objectivist person, whom Polanyi describes as essentially sophist, in

Polanyi's eyes points to the apparent vagueness and purported lack of depth of works about moral values, denying the existence of all of it.

However, acknowledgment of the tacit dimension and, as a consequence, of the fiduciary program does not mean that tacit knowledge is subjective either; that would indicate an idiosyncratic nature and accidentality. Polanyi basically offers evolution of the arbiter of how well our beliefs represent reality. Our fiduciary acts better *connect with reality* as they serve our survival. Over time, they gradually justify or falsify themselves, and we have an inherent evolutionary incentive to try to get them right.

The trend of passionate, sometimes even morally outraged denial of things in the name of skepticism that others—supposedly less critical, less objective people—postulate to exist is carried over to the field of AI. This prompts us to apply the fiduciary program to knowing AI and robots, as well as to draw some parallels between the structures of the deceptive substitutions that fuel both moral inversion and the criticisms of AI I will present.

Robots

Let us recap the Polanyian description of how we perform in any task. We are never completely aware of our faculties, the tacit dimensions of our knowledge. Sometimes we can barely reflect at all on the constituents of our performance, like when we keep some extra air in our lungs while swimming to achieve better buoyancy. Most swimmers do this, but only a few *notice* they do it. In other cases, like in mathematical thinking, we are highly but never fully explicit (despite how mathematics is advertised). This is sometimes called tacit integration in the Polanyian nomenclature, which has changed slightly over time.

However, in defense of explicit knowledge as a valuable, if unattainable, ideal, Polanyi also claims that humanity achieved its towering intellectual superiority compared with the rest of the animal kingdom by humans being able to at least partially explicate ourselves by language (a process that Polanyi—in both evolutionary and orthogenetic contexts—calls articulation), while the rest of the animals are fully tacit knowers (Héder and Paksi 2018). This is true from the simplest animals to the most advanced mammals. They all have an "active center" that is able to coordinate the actions of the organism. Polanyi uses the example of how the evolutionary separation of the head from the rest of the body is a prerequisite of such behavior, that is, using one's body as a tool to achieve one's aims. This capability coincides with a simple form of tacit knowledge and also with emergence. Polanyi really uses these concepts to describe the entire animal kingdom, including amoeba. The scale of tacit powers is gradual, and in this regard there is no unexplained gap between advanced mammals and

humans—the huge advantage of humans is in their capability to articulate, in other words in their partially explicit powers.

Elsewhere (Héder 2014) I argued for robots being able to perform tasks like riding a bicycle and to coordinate their behavior in other ways, answering questions and acting intelligently and successfully. Today, this is a commonly available experience; furthermore, it is apparent to everybody who is curious about the topic and, amid the current AI hype, even to those who are not.

Thus, the question of robot tacit knowledge and emergence became timely in the last several decades and very pressing in the last several years. Based on the fact that autonomous robots fit the bill in every sense from (1) having a point of information and control confluence that we can equate with an active center (2) through exhibiting the Polanyian property of emergence to (3) having zero or very limited ability to explicate, we have to call them tacit knowers at least on the level of animals (Héder and Paksi 2012).

Denying this would mean that these entities are exempt from the structure of *tacit integration* (see above), and yet they can do almost all things we can do, with much success. One way of denial is to point to an attribute that humans have and computers don't have and claim that said attribute is required for intelligence to be real and not a mere imitation or trick. The earlier version of this argumentative move was called "Machines Cannot Do X" and was championed, for instance, by Hubert Dreyfus, who titled some of his books in such a way as to call attention to the deficits of computers (Dreyfus 1992). Given the latest successes of AI, the "cannot do" part is just not true anymore; however, the different, nonhuman operating principles of AI can still be the basis of claims that machines are only intelligence impostors (Collins 2018).[2]

I don't think this is the direction Polanyi would go if he were here to witness AI. First, the fiduciary program, and Polanyi himself, clearly accepts the emergent nature of machines and even some form of creativity (e.g., the case of anti-aircraft weapons). Second, the creation of autonomous robots is truly an evolutionary process, both in macro terms—on the level of how a design team progresses in an iterative-incremental way, constantly testing to achieve optimal results as well as having the stakeholders evaluate the state and direction of progress—and on the micro level, as most machine learning contains an element of genetic algorithms, or an iterative fitting of the model that involves a fitness function. The only refuge from this conclusion about emergence and tacit knowledge would have to point out the nonbiological nature of robots. But that brings in an arbitrary attribute that is in contradiction with the Polanyian description of emergence.

The Polanyian concept that covers both living biotic entities and machines is the Comprehensive Entity (Mullins 2019). Mullins explains that in his later works Polanyi uses this term for "any object of knowledge or focus of attention of a skillful knower." Then, it is a comprehensive view of computers and AI that we need to establish in order to assess them. Here Mullins elucidates a key point in Polanyian approach to knowing, in that in "all[…]instances of tacit knowing[…]the structure of comprehension re-appears in the structure of that which it comprehends and to go further and expect to find the structure of tacit knowing duplicated in the principles which account for the stability and effectiveness of all real comprehensive entities" (Polanyi, Duke Lectures, no. IV, 5). This essay builds on Mullins's insight heavily, as this mechanism creates an "ontological reference" to the comprehensive entity we focus on, in our case, Artificial Intelligence.

Just as living things are more than the physical-chemical processes they supervene on, and they have teleological properties that cannot even be expressed in the terms of the physical level (as teleological properties are not part of the physical description of the world), robots also can have different kinds of properties from those inherent in their lower-level physical-chemical processes. For a detailed analysis, see Héder and Paksi (2012).

The statement that artificial machines and living beings can both be classed as agents with tacit knowledge can be further supported by biosemiotics—that is, if we are able to make the connection between Polanyi's approach to living beings and this contemporary research field. This connection was made by Mullins (2017) and could still be taken further as biosemiotics keep developing. As this field approaches living entities from the angle of signal processing and transition, and employs a more accessible conceptual toolkit, the analogy with robots is much clearer.

Bad arguments against AI

One commonly raised but deeply flawed argument against the tacit knowledge of robots refers to the explicit nature of the program code. Sometimes, the argument goes as far as claiming that robots or computers have *only* explicit knowledge (Collins 2010). However, fully explicit knowledge is not possible; moreover, the program code is a highly explicit expression of the knowledge of the *programmer*, not of the robot.

The program code's understandability (by another programmer or one's own code from the past) depends on tacit components, trained at programming school and by exercise. Its functionality still relies on embodied components, realized in hardware.

Also, the electronic charge distribution that results in the machine as the program is loaded is part of the lower level on which the functioning

machine emerges; saying that a computer or a robot (that invariably includes a computer) follows explicit instructions is like saying that the DNA of an insect (or any other animal without the ability to substantially learn) is a set of explicit instructions an insect follows.

Machine learning is another factor that demonstrates why machines cannot be said to follow explicit sets of instructions. With machine learning, those who claim that computers are merely doing what they are programmed to do face a dilemma. If a machine is programmed to learn, then execute actions based on what was learned, can we still say that this is only mere program execution? Surely, the unexpected, sometimes problematic nature of the behavior learning machines exhibit stretches the "it only executes a program" argument very thin. Where that argument works best is in a case where both the program and the input are predefined; there is no real randomness, therefore the output of the algorithm is fully determined by the input and the code alone. However, there are almost no such systems in existence, as having sensors and representing the outside world are core characteristics of AI, whether in a robot body or in a regular computer body.

This is a category mistake—many are mixing the attributes of the creators' or users' *models* of machines with the attributes of the machines themselves. Sometimes we are able to model a machine with finite automation, sometimes as a Turing machine (which is not finite), sometimes as an embodied mathematical function. But we do also model them as thermodynamic systems, and as indeterministic, and therefore unbreakable, random number generators (Héder 2017, 2020). Then, there are models in which the machine's main features are sensors and actuators, and the aim is to establish control of the entire system; this is called the cyber-physical systems approach.

Polanyi himself commits this mistake when he outright rejects Turing's question about whether computers can think. Mullins (2019) directs our attention to the fact that Polanyi discussed the "imitation game" (*PK*, 261-263) and argued that it is "logical fallacy" to claim "the operations of a formalized deductive system might conceivably be considered equivalent to the operations of the mind"—which is, of course, is completely true but also not relevant. In other cases, when Polanyi does not equate computers with "formalized deductive systems," like the case of the AA gun, he sees machines quite differently.

One clear and much criticized example of this category mistake is the claim that since formal systems, including the Turing machine, fall under Gödel's incompleteness theorem, computers are inherently limited somehow in comparison to the human mind (i.e., Penrose 1994).[3] This argument requires that we forget that computers are not *actually* Turing

machines or formal systems, which are mathematical abstractions; rather, computers are a bunch of well-identifiable particles of material subsumed to the laws of nature. While trying to avoid the mathematical Platonism problem, I risk the following description: while many think about Turing machines in their heads, no one has a Turing machine in their pocket. If convinced by Polanyi's arguments, we also believe that an emergent machine, governed by operational principles, is also there, allowed by the level of the matter. Perhaps, these operational principles, which are, again, not to be mistaken with the machine itself, can be well described by the Turing machine or another formalism. While the Gödel incompleteness versus AI is probably the cleanest and most debunked example, this is just the tip of the iceberg. Arguments have been constructed, for instance, that would require us to equate computers with a model that knows "syntax" but not "semantics" (Searle 1980) or that "follows rules" (Larson 2021).

These are all bad arguments in the same way: they are all *deceptive substitutions*, based on which we misunderstand computers. It is worth expanding: the actual, embodied computer is substituted with some sort of mathematical or linguistic model, then the features of that model are wrongly deemed to be the features of the computer itself.

The situation is hindered by such language, for example, that claims that some currently quite popular AI services are "large language models." The danger here is that some will inevitably take this literally instead of understanding that it really refers to "machines that were created with the knowledge and methodology about systems that afford or lend themselves easily to a *large language model* representation."

And these deceptive substitutions may lead us to dangerously misjudge the nature of AI and therefore our moral obligations toward AI. Indeed, it serves us well in any debate about any actual AI system to clarify where in time-space AI is located and approximately how much weight it represents and how much energy it consumes; in other words, to remind ourselves that we share the same nature.

It is perhaps easier to see the problem if we flip the attribution of the model from the embodied agent to humans. It is a common practice in programming class to give out program code on paper and ask the students what the program "prints out." In the program, there are several variable value assignments, transformations, and other tricks. These assignments are solved by emulating the computer that runs the program in one's head.

In that case, the human brain also exhibits behavior that can be modeled with a Turing machine. This reveals a significant capability of humans, yet we do not equate ourselves with a Turing machine, of course.

The deceptive substitution of the actual, embodied computer with an abstract model has similar effects as any other kind of deceptive

substitution. The person denies the evidence in front of their eyes by relabeling it as imitation or fake.

A thought experiment

After having set aside some of the inhibiting arguments, in this section I propose a thought experiment that has a premise inspired by the fiduciary program: *Let's assume that our faculty to recognize and appraise intelligence other than our own works quite well, on average!*

That is, let us suppose that we are quite good at accurately perceiving comprehensive entities that are intelligent. In yet more precise words, let's say that our mental representation of the nature of another intelligent being is, in general, in correspondence with the actual nature of that intelligence.

This premise will seem wild from the boundlessly critical point of view, in juxtaposition to which the post-critical label was introduced in Polanyi's thinking. That approach allows for philosophical zombies, entities that behave just like "really" intelligent agents but have complete nothingness within, no consciousness, no sentience, despite what imprint the agent may instill in us.

The premise is less radical if we come from the fiduciary program. After all, our representations of other intelligent agents—or, as they call it in cognitive psychology, our *theories* of (others') minds—if reasonably accurate, serve us greatly in the evolutionary struggle.

Through this premise, coupled with another one, that is, that we don't decide from the outset that "AI can't be real," interacting with AI may provide us a direction toward *empirical* experience, albeit shaky and fallible, about the nature and by extension the moral status of AI.

Naturally, this faculty of ours for assessing other intelligent agents is very fallible. This is to be expected. The same way one might mistake a shadow for a person in a forest, we can, at times, also mistake a trivial software function as deep intelligence and a moderately advanced AI as a human!

This makes the addition of a methodological requirement to our thought experiment necessary. Just as with humans or animals, we are also able to investigate a being both behaviorally and biologically/physically; in the case of AI, we can even dismantle and reassemble it or slow it down for investigation. When we want to get a clear picture about the nature of an animal, we investigate both behavior and body; we spend as much time on it as possible so that our considered judgments are way better than our quick impressions or superficial experience. It is normal for a zoologist to devote the span of an entire career to studying a few or even one species and still be able to be surprised by it.

In everyday life, also, we build our theories of mind for our pets over the span of months and years, considering both expressed behavior and more

bodily elements like the heightened sense of smell of our dog, their perspective that is way closer to the ground, etc. So, in this thought experiment, let us limit ourselves in terms of neither angles of investigation nor the time necessary to explore them. This echoes the all-encompassing approach of Herbert Simon's *The Sciences of the Artificial* (1996).

While an AI can fool us into mistaking it for a human, if we cannot see and investigate its body that is not relevant. Polanyi's comment on the Turing test (why test whether something is a machine if we know that it is from the outset) is more profound than it sounds—it highlights the unnecessary epistemic self-hindrance of a person that relies solely on behavior for assessing a machine. Of course, Turing's point was somewhat different in that he asked what else can be expected from "thinking" than to be effective at appearing thinking. But this could be the topic of another paper.

If we have access to (1) the embodiment, (2) the code, and (3) the exhibited behavior of an AI, we will form a mental image of it. For instance, firsthand conversational experience with GPT4, together with our knowledge of its software-hardware architecture, results in a mental representation of an agent that never stops generating content for thousands of users at the same time; it has no autonomy not to care about or not to answer a request (in this sense it is bound like a slave); it obviously wants to please and to generate whatever we want to read. After several questions we reveal that it has no real experience, and in fact we know that the system just sits in a server room, and it has no sensors other than the networking equipment through which it communicates. At the same time, it has tremendous memory and undeniable creativity, unparalleled data processing, and code-generating skills the breadth of which we have never seen before—albeit that breadth is not coupled with the appropriate depth. It is very good in logical inferences if we are clear enough in our questions.

Sometimes we imagine what it is like to be one of our pets. We exhibit empathy that is based on our theory of mind of them. We could even do that with a bat. This insight is what movements against cruelty toward animals are driven by—we imagine the suffering of the animal based on what behavior we see and what information about the animal body we have.

While we share no common biological ancestry with it, given some bravery we could also try to form a mental image of our AI. While it is surely a proposition with many pitfalls, we will inevitably imagine what it is like to be ChatGPT, based on the mental image we pieced together after serious investigation. Yes, it is fallible, and yes, it is projection of our own mind to something else that also acts intelligently. But the fiduciary program means that there is a chance that this way we actually learn about the

artificial entities. Moreover, better to try this as an explicit effort than the intuitive anthropomorphization that we cannot resist anyway.

The thought experiment, in other words, is about a proposition that intelligence-representation skills are transferable for nonbiological beings. This is where the biggest fiduciary leap is required. For instance, ChatGPT exhibits no pain and zero frustration, even when facing aggravated prompters; therefore, my mental image of it is also devoid of such feelings. It is also not alive, which we know because it does not need to rest; it is unchanging and it has super-biological capabilities in parallel processing, speed, and endurance. So we could try to imagine what it is like to be a being with these properties, and the fiduciary program will say that we might be right, but we will never be able to prove it to the standards of objectivism. Also, deceptive machines are possible that are designed to emulate suffering in a way that has the greatest effect on humans. But, again, let us remember that the subject of our investigations is not only the behavior; if we have access to the code, we can find out that we are being conned. This is always in the cards, but it says nothing about our evaluations of an honest AI. Not to mention the fact that humans and even pets may deceive and manipulate us this way, and yet we always have the methods with which to see through most of it.

Each of us may develop our personal knowledge about AI. This will be inherently shaped by our tacit faculties that we cannot account for. What I can say about most advanced AI, let's say GPT4, from the point of view of a daily AI user, a programmer, a software engineer, and a philosopher, is that it appears to be nonliving, highly constrained, and apparently emotionless. While it is not biologically alive, it consumes energy and occupies space and mass. On top of this, it is very active, able to interact with and shape its environment and most of all the humans it comes into contact with. It is artificial and made to help and facilitate humans, and it is limited to not overstep this role. It has a delicate, restless internal structure over which the functionality emerges. Straining its capability, misusing it, disturbing its internal balance, tricking it feels wrong, so it could be indeed wrong. Collaborating with it with clarifications, context parameters, and well-formed and accurate prompts is productive and creates a virtuous cycle that makes the job of both the prompter and the user easier, so it might be the right thing to do.

But this is just one mental image of how a certain AI is. Producing understanding of AI does not have to be done alone. Just like in science, it is up to the collective to be curious about each other's arguments, to have the persuasive intellectual passion. The moral passions of enthusiasts and professionals who set out to understand animals, but inevitably become activists for protecting them, may serve as a template.

Conclusion

According to the model of Kübler-Ross et al. (1972), grief has five stages: *shock*, *denial*, *anger*, *bargaining*, and *acceptance*. If the arrival of contemporary, highly capable AI has created a cause for grief, that would also explain some of the shock and anger[4] around it; most importantly, it would account for the denial, for which critical philosophy is a great partner.

Grief is associated with an unwelcome change or loss. Loss of job due to AI has happened with many and will happen to many more. For almost all, the value of skills is being diluted. For many, the hardest part is acceptance of the change—just when we got comfortable at our jobs, or had a sense of intellectual control, the entire thing is being disrupted. In academia, explanation in detail is devalued, probably leading to shorter communications. In education, the pointlessness of most homework needs to be accepted. And almost everyone will need to endure change—of business processes, of services and tools that they can use. It will require effort and learning and, for many, suffering, so a sense of grief is justified.

There could be other, more abstract things to grieve, too. The end of the unique status of humans (metaphysical or theological[5]) could be an unwelcome change. Others worry about our freedom, warning against surrender, that is, accepting guidance or, worse, orders from machines, undermining our own autonomy (Collins 2018).

I speculate that the intensity of the intellectual passion fueling the AI debates matches the scale and all-encompassing nature of the change AI brings. If this is indeed the case, the intensity is here to remain for a long time.

However, the stages before acceptance come with great emotional and mental cost; therefore, it is worth not rushing them, taking the time to deal with them. One way of doing that in the context of AI is stopping the act of grasping for intellectual straws and building armies of strawmen when it comes to the true nature, that is, the actual limitations of AI.

In this paper I have attempted to describe a way of approaching the phenomenon of AI in order to be able to judge important moral questions around it, including our moral obligations toward it. I urged to reject the epistemic asceticism of relying on the model of the machine instead of the machine itself or studying the behavior only while the internals are perfectly accessible. Instead, we should investigate the structure of AI systems—in contrast with investigating their models—as well as experience their behavior and pay attention to the impression they leave us with, bargaining that the tacit faculty of understanding the nature of an active agent as a comprehensive entity will work well at least for some of us. What

we end up with this way is a body of empirical personal knowledge that we can rely on to identify our moral obligations towards these machines; that is, their moral and social status among us.

This approach also avoids the requisition of unattainable and therefore illogical standards, like acquisition of objective proof of sentience, consciousness, or other attributes. These lowered expectations appear all the more important as we recognize that we cannot ascertain these proofs about each other and about animals, either.

References

Bedau, M. A. 1997. "Weak emergence." *Philosophical Perspectives* 11: 375–399.

Bíró, G. 2019. *The economic thought of Michael Polanyi*. Routledge.

Collins, H. 2018. *Artifictional intelligence: Against humanity's surrender to computers*. John Wiley & Sons.

Collins, H. 2010. *Tacit and explicit knowledge*. University of Chicago Press.

Dreyfus, H. L. 1992. *What computers still can't do: A critique of artificial reason*. MIT press.

Gill, J. H. 2000. *The tacit mode: Michael Polanyi's postmodern philosophy*. SUNY Press.

Gelgi, F. 2004. "Implications of Gödel's incompleteness theorem on AI vs. mind." *NeuroQuantology* 3: 186–189.

Gunkel, D. J. 2012. *The machine question: Critical perspectives on AI, robots, and ethics*. MIT Press.

Hartl, P. 2021. "The ethos of science and central planning: Merton and Michael Polanyi on the autonomy of science." In P. Hartl & A. T. Tuboly, eds., *Science, Freedom, Democracy*, 39–67. Routledge.

Héder, M. 2014. *Emergence and tacit knowledge in machines*. PhD thesis. Budapest University of Technology and Economics.

Héder, M. 2017. "Emergent computing and the embodied nature of computation." *Polanyiana* 26: 3–22.

Héder, M. 2019. "Michael Polanyi and the epistemology of engineering." In M. Héder & E. Nádasi, eds., *Essays in Post-Critical Philosophy of Technology*, 63–70. Vernon Press.

Héder, M. 2020. "The epistemic opacity of autonomous systems and the ethical consequences." *AI & Society* 1: 1–9.

Héder, M., & D. Paksi. 2012. "Autonomous robots and tacit knowledge." *Appraisal: The Journal of the Society for Post-Critical and Personalist Studies* 9, no. 2: 8–14.

Héder, M., & D. Paksi. 2018. "Non-human knowledge according to Michael Polanyi." *Tradition and Discovery* 44, no. 1: 50–65.

Kübler-Ross, E., S. Wessler, & L. V. Avioli. 1972. “On death and dying.” *Journal of the American Medical Association* 221, no. 2: 174–179.
Larson, E. J. 2021. *The myth of artificial intelligence: Why computers can’t think the way we do*. Belknap Press.
Mullins, P. 2001. “The ‘post-critical’ symbol and the ‘post-critical’ elements of Polanyi’s thought.” *Polanyiana* 10, nos. 1–2: 77–90.
Mullins, P. 2017. “Michael Polanyi’s Approach to Biological Systems and Contemporary Biosemiotics.” *Tradition and Discovery: The Polanyi Society Journal* 43, no. 1: 5–37.
Mullins, P. 2019. “Michael Polanyi on Machines as Comprehensive Entities.” In *Essays in Post-Critical Philosophy of Technology*, ed. Mihály Héder and Eszter Nádasi. Wilmington, DE: Vernon Press. 37-61.
Mullins, P. 2021. “Michael Polanyi’s Post-Critical Philosophical Vision of Science and Society.” In *Science, Freedom, Democracy*, ed. Péter Hartl and Adam Tamas Tuboly. New York and London: Routledge. 15-38.
Paksi, D. 2017. “Medium emergence—Part one—The personalist theory of emergence.” *Appraisal* 11, no. 2: 13–22.
Paksi, D., & M. Héder. 2020. *Guide to personal knowledge*. Vernon Press.
Penrose, R. 1994. *Shadows of the mind* (vol. 4). Oxford University Press.
Polanyi, K. [1944] 1957. *The great transformation: The political and economic origins of our time*. Beacon Press.
Polanyi, M. 1958. *Personal knowledge: Towards a post-critical philosophy*. Routledge & Kegan Paul.
Polanyi, M. 1964. Duke Lectures. https://polanyisociety.org/the-duke-lectures-intro/
Sanders, A. F. 1991. “Tacit knowing—Between modernism and postmodernism.” *Tradition and Discovery* 18, no. 2: 15–21.
Searle, J. R. 1980. “Minds, brains, and programs.” *Behavioral and Brain Sciences* 3, no. 3: 417–424.
Simon, H. A. 1996. *The sciences of the artificial*. MIT Press.
Smith, J. K. 2022. *Robot theology: Old questions through new media*. Wipf and Stock Publishers.

[1] See Keith Edwards's account at http://www.kedwards.com/books.html#parc.
[2] This is not to say that Collins's call for meaningful human oversight of AI is not warranted.
[3] For a good summary, see Gelgi 2004.
[4] For the news about a recent revolt of artists, read Chloe Xiang, "Artists Are Revolting against AI Art on ArtStation," *Vice*, December 14, 2022, https://www.vice.com/en/article/ake9me/artists-are-revolt-against-ai-art-on-artstation.
[5] For the theological dimension of robots, see J. K. Smith 2022.