# The Metonymical Trap

*Éloïse Boisseau – Aix-Marseille Université – draft – the final version of this chapter can be found in Helliwell, A. C., Rossi, A., & Ball, B. (eds.),* Wittgenstein and Artificial Intelligence*,* Vol. 1: Mind and Language*, Anthem Studies in Wittgenstein, Anthem Press, pp. 85-104 (2024).*

## From the Mereological Fallacy to the Metonymical Trap

Maxwell Bennett and Peter Hacker (2023) famously introduced the so-called charge of 'mereological fallacy'. This fallacy is not per se a *reasoning* fallacy but is more akin to an *attribution* mistake (an attribution mistake that can admittedly lead to fallacies of reasoning). In the context that is of particular interest to Bennett and Hacker's discussion, the 'mereological fallacy' is specifically related to neuroscience:[1] to say, for example as part of a neuroscientific explanation, that the brain *constructs hypotheses* would be a paradigmatic instance of a mereological fallacy (*ibid*, 80). A brain is indeed nothing more than a part of a human being, and only of a whole human being does it make sense to say that it constructs hypotheses.[2] Therefore, transferring the qualities and capacities of the whole (the human being) to the part (its brain) would not be a viable philosophical option. It would not be an option in particular because it is partly *constitutive of the meaning* of the attributed expression ('to construct hypotheses') that a range of behaviour be associated with it, a range of behaviour that a brain obviously lacks. This idea is Wittgensteinian through and through, and found in the oft-quoted passage of the *Philosophical Investigations* (Wittgenstein 2009: PI §103): 'It comes to this: only of a living human being and what resembles (behaves like) a living human being can one say: it has sensations; it sees; is blind; hears; is deaf; is conscious or unconscious'. A brain does not – in any way, shape or form – behave like a living human being.[3]

---

[1] In other contexts, the mereological fallacy essentially boils down to two general attribution mistakes that logicians sometimes refer to as the 'fallacy of composition' (a part can have some properties that are not transferable to the whole that it is a part of – e.g. parts of this machine might be little while the whole machine might not be) and the 'fallacy of division' (a whole can have some properties that are not transferable to its parts – e.g. this machine is large while its parts might not be). Cf. for example Walton (2008, 156 *sqq*).

[2] One might also argue (and I thank an anonymous reviewer for the suggestion) that the verb 'construct' in the expression 'to construct hypotheses', should in any case, be read as metaphorical (or analogical). One does not *literally* construct hypotheses (hypotheses are not the kind of things that can be constructed) but one does literally *formulate* hypotheses. I am leaning more towards the idea that this first expression can be (and is) taken literally (there is no category mistake at play in saying that someone has constructed a curious hypothesis). This is precisely an instance of what George Lakoff and Mark Johnson (2003, 51) qualify as a 'literal expression structured by metaphorical concepts'. Our verb 'construct' for the formulation of hypotheses is a literal expression: saying that we *construct* a hypothesis is indeed an ordinary way of describing what we do when we formulate hypotheses – even though this way of speaking is indeed metaphorically structured by, say, the model of a *physical* construction.

[3] One of the central insights of Wittgenstein on this issue consists in pointing out that one condition for saying that someone is capable of doing a particular thing – in our present case: formulating hypotheses – is that we understand what it would mean for such an individual to *express* or *display* this kind of capacity in their behaviour (i.e. in their words, gestures and actions). This first condition is closely related and in fact goes hand in hand with a second necessary condition for saying of an individual that they are endowed with an intellectual or mental capacity, which is that we understand the role this capacity might play in the individual's life. In this respect, one can see why a brain might be a

Returning to the more general diagnosis of mereological fallacy: although there are numerous instances of predicates that can equally and harmlessly be attributed indistinctly to a whole or to a part of that whole (the fallacy is not systematic) and even if mereological fallacies can apply in a variety of contexts (to borrow one of Hacker's examples (2013, 287) a clock indicates time, a capacity that its fusee or face could not be said to possess), we can say that such a diagnosis is particularly fruitful when it concerns the so-called psychological predicates. These predicates relate to the mental life and the cognitive or psychological capacities of a person. They are often contrasted with predicates that are merely physical. Saying that a stone is heavy uses thus a physical predicate, when on the other hand saying that Alfred wants to buy bread, that he is in pain or in a hurry to get home, uses various psychological predicates (as all these relate one way or another to Alfred's mental life).[4]

What is of particular interest for our present purpose is to remark that the usual way out, or the sort of 'loophole' for those accused of committing a mereological mistake is to understand these ascriptions in a *metonymic* way or 'metonymically' (this way out is mentioned by Bennett and Hacker (2023, 83)). How is this a way out? First, what is a metonymy? A metonymy is a figure of speech that uses a salient aspect of a thing to refer to the thing itself (or possibly to refer to another thing that is intimately related to it) (Littlemore 2015). In the confines of this chapter, I emphasize that it is the only dimension that I will retain from the sometimes contradictory taxonomies of metonymy: I take it that metonymy has to do with *identifying* the subject of a statement – and I will furthermore say of the identification that it is *indirect* when metonymy is involved. Metonymy is thus characterized by this *referential* aspect.

I would like to suggest that the mereological fallacy is a special case of a more general fallacy: the metonymical fallacy. The mereological fallacy has to do only with the part-whole relation, while the

---

poor candidate for being the subject of cognitive abilities. Observing a brain in activity might mean observing the transport of oxygen via the blood from one region to another or observing the firing of nerve impulses from the neurons. It thus means observing no behaviour sufficiently similar to that used as a criterion for the usual attribution of the activity of formulating hypotheses. Although oxygen variations in the brain may well be necessary events to *enable* the individual whose brain it is to formulate hypotheses, these variations are not, in themselves, part of what it means to formulate hypotheses (they are not constitutive of what it is to formulate hypotheses). Furthermore, a brain is not a whole individual with a complex life in which the activity of formulating hypotheses would be relevant. A brain, like a pencil or a lung, is therefore not the kind of thing with a behavioural repertoire that makes it possible for it to manifest such a faculty. In the same way that Wittgenstein (Wittgenstein 1980: RPP §192) once remarked how we would fail to grasp what it would be like for a table to think, it seems that we also fail to grasp what it would be like if a brain were to think, construct hypotheses and so forth: 'But if I say "A table does not think", then that is not similar to a statement like "A table doesn't grow". I shouldn't know "what it would be like if" a table were to think. And here, there is obviously a gradual transition to the case of human beings.'

[4] Wittgenstein frequently focused on the question of the status of psychological concepts (Wittgenstein 1980: RPP *passim*) and has notably (e.g. Wittgenstein 1980: RPP §63 and Wittgenstein 1958: BB, 66–67) sketched out – rather tentatively – an important logical or grammatical asymmetry in psychological predicates when used in the first person of the present (on the one hand) and in the second or third person of the present (on the other hand): it is arguable that when I say 'I am in pain', 'I believe *x*', 'I wish I had done that', and so on, I usually do not intend to describe anyone (not even myself) whereas it is exactly what I intend to do when I say 'she is in pain', 'she believes *x*', 'she wishes she had done that'. Used with the first person (in the present tense), these psychological expressions are better thought not as ways *to describe something* but as ways to *express something*. In this respect, the possibility of error is non-existent (and as such there is no *identification* in play). As for physical concepts, they have (in the present tense) exactly the same purpose in an 'I-statement' and, say, in a 'she-statement'. Thus saying 'my hair is black' and 'her hair is black' both require some kind of observation, and I am in both cases describing someone that I have identified.

metonymical fallacy involves a more general feature-thing relation (a part of a thing can be considered as an aspect or a feature of a thing). Both can lead to what I will call the metonymical trap (taking an indirect identification for a direct one). This focus on the more general fallacy will allow us to go beyond the strictly neuroscientific debates discussed by Bennett and Hacker (2023) while keeping the results of some of their analyses. I would like to suggest that it may be philosophically fruitful to bear in mind the risk of a metonymical trap when considering the status of some attributions to machines. My aim in this chapter is to discuss and evaluate the question of the attribution of predicates to machines (in particular cognitive or intellectual predicates).[5]

## Metonymy and Artefacts

Among the cases of apparent attributions of intentional characteristics to machines, some leave no room for doubt as to their metonymic nature. We can thus easily distinguish between the metonymical statement:

*the trains are on strike*

and this other non-metonymical statement (although both are syntactically on a par):

*the trains are out of order*

Whereas in the second sentence something is indeed said about the trains themselves (i.e. that they are not working, that there is a technical issue *with them*), this is obviously not the case in the first proposition, in which if anything is said about the trains, it is only said in a derivative or consequential manner (e.g. it can be inferred – if it is true that the trains are on strike – that not a single train is currently in circulation). Unlike the second statement, the first one then does not literally refer to (or talk about) the trains (and so nothing is ascribed to them); the reference is thus *secondary*, *oblique* or *derivative* because it is obviously not the trains themselves that are on strike (demonstrating, protesting, etc.), but the drivers of these trains (it is they, not the trains, who have this range of behaviour at their disposal). There is thus an asymmetry that justifies the idea that the reference to trains is secondary here: you have to under- stand what it means for *people* to be on strike to understand that 'the trains are on strike', but you do not have to understand what it means for *trains* to be on strike to understand that 'people are on strike'. This first statement has therefore a metonymic character since we seem to attribute to the machines (in this instance, trains) a behaviour (being on strike) which is in fact that of the agents who operate them. In this particular case, the figure of speech involves referring to the train drivers by means of the machines they operate. If we bear in mind, as we mentioned earlier, that metonymy works by an aspect-thing relation, the train drivers are then the thing ultimately referred to, and the aspect related to them through which the reference obtains is the trains they operate. This type of metonymy is sometimes referred to as the 'object used for user' metonymy (Lakoff and Johnson 2003, 38). The figure of speech occurs with a shift in the subject of reference: it is indeed the drivers we are referring to, although we do not mention them directly.

[5] A machine is traditionally an artefact with moving parts. We will follow the more recent use (prevalent in AI circles) and speak of 'machine' to cover both the case of software and that of robots. What is common to these two types of machines is that they are both contingent and dependent artefacts: abstract artefacts in one instance, concrete artefacts in the other. On the typology of abstract artefacts, see Amie Thomasson (1998).

The predicate of the sentence, on the other hand, should be taken quite literally. It is indeed the very act of being on strike that we want to state in the proposition. 'Being on strike' is not, here, used indirectly or obliquely as a predicate; it is not some sort of proxy for *another* predicate: we are not in fact intending to use the phrase to refer to any *other* activity than that of being on strike. If this were the case, we might then be leaving the realm of metonymy and entering that of metaphor: in broad strokes, we could suggest that metonymy has to do with an indirect identification of a subject, while metaphor has more to do with an indirect or tangent predication. In this regard, Lakoff and Johnson (2003, 6) argue that 'the essence of metaphor is understanding and experiencing one kind of thing in terms of another'. The point of a metaphor would be to use the lexical field from a specific concept in the context of another concept. The example examined at length by Lakoff and Johnson is the 'argument is war' metaphor. They point out that the activity of debating is often understood and referred to in terms of the activity of fighting (waging war): thus, we speak of '*defending* a position', '*attacking* an argument', of 'an argumentative *strategy*' and so on. These terms arise from a primary conceptual field (that of war) and are then transferred (applied) to another conceptual field (that of debating). This is not what is at stake with metonymy since – as we have seen – it is not *what is said* that is on the line: it is what it is said *of*.

Going back to metonymy, let us note that this very common figure of speech is, at first sight, quite harmless. It is harmless as long as one keeps in mind that what is in play is only a secondary, derivative use or attribution and that it is in no way intended to bestow human characteristics on the machine. Therefore, if taking the proposition literally is problematic, the problem does not lie in the misidentification of the predicate, but in the identification of the subject. As we have just discussed, trains are not the kinds of things that can go on strike: the grammatical subject of the proposition is not suited (when taken literally) for the predicate at issue. We shall then distinguish between the literal subject (when what I seem to be talking about is indeed what can receive what I say about it) and the metonymic subject (when what I seem to be talking about is *not* what can receive what I say about it).

Furthermore, in the 'trains are on strike' example, the metonymic component is very easy to detect, and in this case nobody would likely fall into the metonymical trap. In other cases, however, this component might be more difficult to identify. What should we think, for instance, of the apparent attribution of a cognitive faculty to a machine, as in 'the pocket calculator calculates'? I would like to argue that in this case – just like in the 'trains are on strike' case – the machine (the pocket calculator) is *not* the literal subject of the action but only a metonymical one. More generally, I would like to argue that the attribution of cognitive or psychological characteristics or capacities to a machine can never be literal but can only be metonymic. This echoes a remark made by Wittgenstein (Wittgenstein 2009: PI §103) where he notes that the attribution of psychological predicates to inanimate things (he takes the example of children attributing pain to dolls) can only be secondary uses of the terms at play. I will come back to this. For the moment let me simply note that the claim that a machine can only ever be a metonymical subject of attribution of cognitive or psychological capacities is, of course, much more controversial than merely saying that there is a metonymy involved in the sentence 'the trains are on strike'. That is why it will be useful to distinguish between two broad classes of attribution of actions to artefacts (in particular machines). In the next section, I will begin by examining the status of the attribution of actions that might

simply be described as 'physical' or 'natural'. The question is whether – in propositions such as 'a plane flies', 'a blender grinds', 'the car is moving' and so on – these attributions are literal or metonymic. I will argue that they are indeed literal, that is, that the machine is genuinely the subject of these actions. I will then examine the question of the attribution of what might be described as 'intentional', 'intellectual' or 'cognitive' actions. My concern will be to determine whether, in statements such as 'this pocket calculator *calculates*', 'the software *translates*' or 'this recognition network *discriminates*', and so on, the subject of the attribution is being identified directly or metonymically. In this instance – inspired by Wittgensteinian insights – I will attempt to show that, contrary to the previous case of the attribution of 'physical' or 'natural' actions, the attribution of cognitive or intentional actions to machines is necessarily metonymic.

## The Machine as a Literal Subject of Attribution of Non-Intellectual Actions

Since it is always possible to ask, when presented with a machine, what it actually does,[6] a distinction should be made between two broad classes of answers: answers that call for natural or physical descriptions ('the fan blows air', 'the vacuum cleaner sucks up dust', etc.) and answers that call for more intellectual, cognitive or intentional descriptions ('the computer calculates', 'the software discerns characters', 'the model learns from its past mistakes', etc.). The first kind of cases of ascription of actions to machines that we will consider are cases of non-intellectual actions. By 'non-intellectual actions', I mean actions that are usually considered as not necessarily requiring any intelligence to be performed. In contrast to actions that would instead be intellectual (such as speaking, reading, calculating, translating), these actions are likely to be easily performed in a purely mechanical way. The action of the toaster to toast or burn my toast, the action of the crane to lift the cinder block or the action of the plane to fly are thus examples of actions that could be classified as *non-intellectual*.[7]

Let us focus on the example of 'the plane flies'. In this sentence, is the subject of the action identified directly (as it was previously the case in the proposition 'the trains are out of order') or metonymically (as it was in the proposition 'the trains are on strike')? Unlike 'being on strike' – which is a specifically human and intentional action and can therefore only take place within a particular form of life (made of norms of organized labour, uses, practices and institutions) – it is

[6] Even if the answer is 'nothing'! It is indeed necessary to make this distinction: for any mechanical artefact or for any machine, it is always possible to ask: What does it do? However, it is not true that it is possible to ask this for any *artefact* (*sans phrase*). It would be odd, for instance, to ask what a sheet of paper (a full-blown artefact) does. Not because – like the famous 'useless box' (a full- blown machine) – it does nothing (meaning nothing interesting, nothing useful, etc.) but because the question is simply incongruous (misleading): a sheet of paper is not the kind of thing that does (or does not do) anything at all. All artefacts have an end (one can always ask: 'What is it for?'), but it is not true that all artefacts have a typical action or range of actions.

[7] As an anonymous reviewer remarks, one *could* resist this distinction between intellectual actions and non-intellectual actions by adopting a physicalist stance, thus maintaining that *all* actions are ultimately non-intellectual. Since I hope to show that in the case of the attribution of actions to machines, there *is* a discrepancy at play (as to whether the attribution is literal or metonymic), this discrepancy could ultimately give us good reason to recognize that the distinction in question is relevant.

not obvious that the action of flying should be dependent on the wider framework of a form of life governed by uses and norms (though there are of course requirements for the action to even take place: there can be no flight if there is no atmosphere, etc.). It is therefore not unusual to ascribe this action of flying not only to beings whose existence is not governed by norms and institutions (birds, moths, flies, mosquitoes), but equally to inanimate substances (be they natural or artefactual). We can thus easily say that *leaves*, *paper planes* and *dust balls* fly. Besides, there is nothing metonymic about these examples: if there were, we should be able to explain the figure of speech involved (and in particular we should be able to identify the proper referent). In other words, if these subjects were metonymic subjects of action, we should be able to identify what they really refer to. Yet, there is no obvious indication that the subject 'leaf' in the proposition 'the leaf is flying' is used to refer to a subject of the action of flying *that is not* the leaf. Similarly, there is no reason to believe that the predicate 'flies' – in this same sentence – is used with the literary purpose of representing an action *other* than the action of flying (and it is hence doubtful that it might be metaphoric).

These examples can also be contrasted with other examples which are definitely metonymic. Saying that this *plane* flies from Paris to London does not have the same obvious metonymic weight as saying that this *airline* flies from Paris to London. This second proposition has a clear metonymic dimension: whatever the verdict may be in the case of the plane, it seems clear enough that airlines are not the kinds of things that fly from one point to another. Our first impression is therefore that the attribution of physical characteristics to a machine should be considered literal (non-metonymic). Although we shall ultimately endorse this conclusion, let us nevertheless see what concerns one might express with regard to it – and how these can be met.

## The Powers of Machines

Going back to 'the plane flies', there is a vexing difficulty, operating sideways, that we must delve into at this point: one can feel uneasy with saying that a plane flies *in the same sense* as one says that a bird flies. Why is that? Because it seems that we are ascribing to the former an action which ultimately does not depend on itself (upon which – contrary to the latter – it lacks a form of control and spontaneity). The plane – like the Frisbee or the kite (but unlike the bird) – only flies because *we* (i.e. external agents) have decided so (or because we have decided to let it fly or because we have built it so that it can fly, etc.). Thus, there is a noticeable discrepancy between the bird's power to fly and that of the plane: the bird can actualize its power on its own, whereas the plane cannot (and requires the intervention of an external agent – namely, a human being).[8] It might then be tempting,

[8] This consideration also applies in the case of machines that are often misleadingly described as 'autonomous'. In the industrial sector, the term 'autonomous' applies to artefacts that do not – at all times – require human assistance to perform the tasks for which they have been designed. There is assuredly a *continuum* in what should count as 'autonomous': if I voluntarily drop a hammer on a nail, it is in a way, while free falling, 'autonomous', but the range of the outcome is rather limited. A military drone, on the other hand, does not require the constant supervision of a human being to fly and the range of the outcome can be very diverse (it is this wide variety of outcomes that calls for these artefacts to be given a proper legal framework). Of course, the fact that the drone *can* fly necessarily depends on the prior intervention of human beings (who created it, who have a use for it, etc.). The drone can only exercise its power within the confines of its programming – which, admittedly, may be quite broad. This programming has nonetheless been designed and set by human makers. And most importantly, as we will see in the next subsection, a drone does not have a proper *two-way* power.

for our present purposes, to con sider that in a way the plane does not fly *at all*, but that *we* fly *with it* or *by means of it*. There would be a twist here: the plane would only be a simple metonymic subject of the action of *flying*, the real subject being the pilot who uses it to fly (as an instrument). Does this mean that the plane should be denied *any kind* of power or action?

## *One-way versus two-way powers*

An old but useful Aristotelian distinction recently brought up by Peter Hacker (2010) might help us clarify this vexing question. It is the distinction between a *one-way* power and a *two-way* power.[9] To have a one-way power, Hacker explains, is to have a capacity which is actualized whenever the circumstances necessary for it arise. An agent endowed with a one-way power will therefore exercise its power whenever these conditions are met. In contrast, to have a two-way power is not only to be able to actualize one's capacity whenever the circumstances are favourable, but it is also to be able *not* to actualize it. The plane's power to fly would thus be a one-way power, since when all the conditions are met for the power to be actualized, the power will in fact be actualized. The bird's power, on the other hand, would be a two-way power, since even when all the relevant necessary conditions are met, the bird can either exercise its power or refrain from doing so. Following Hacker (2010, 95), we can then say that the circumstances that are necessary for the plane to exercise its power are *occasions* for it to exercise its ability to fly, whereas the circumstances necessary for the exercise of the bird's power are *opportunities* for it to exercise its ability to fly (see also Alvarez, 2013). This distinction makes it thus possible to separate two types of action which correspond to these two types of power. It is now clear that the plane does not act *in a sense that it actualizes a two-way power*. Still, there is a definite action of flying when the plane is in the air, and it must be said that this action involves it *directly* (and not metonymically). Thus, even though machines can literally be said to have and exercise a given power, this power in itself requires further clarification – a clarification that we partly find by means of the distinction between one-way and two-way powers. We shall now continue to pursue this clarification.

## *Mere Cambridge agents*

I would like to tentatively introduce another distinction that might prove useful for my purpose. It is a distinction parallel to the one Peter Geach (1969, 71) drew regarding changes when he coined the phrase 'mere Cambridge changes'. First, I will remind the reader what is at stake with Geach's terminology.

Suppose a thing $x$ changes when a predicate that was truly ascribable of $x$ at a given time is no longer true of $x$ at a later time (or conversely, if a predicate that was not truly ascribable of $x$ at a

[9] The distinction between one-way and two-way powers can be traced back to Aristotle, whose point was essentially stated in terms of a distinction between *rational* and *non-rational* powers. A rational power, according to Aristotle, is 'a capacity for contraries' (2016*,* 1046b5 and 1048a5- 20), while a non-rational power is not. See also Alvarez (2013).

given time, then becomes true of *x* at a later time).[10] This would be a purely logical conception of change, which has only to do with a difference between the truth value of a proposition at a time and its truth value at a later time. This way of thinking about change is what Geach refers to as 'Cambridge changes' – and this conception used as a criterion for saying that a thing changes is what he labels the 'Cambridge criterion'. However, Geach remarks, we want to distinguish, within this broad logical conception of change, changes *that really affect the subject of change* and those that do not. On the one hand, then, we have *real* changes, that is, changes which happen to, and really affect, a given subject: for instance, a wall may have a change in colour (initially yellow, it becomes blue after being repainted). Someone may once have had hair, but no longer does (thus changing from hairy to bald) and so on. These are changes *of* the subject of change. It is usually this type of change – *real* change – that we think of when we ordinarily speak of 'change'. On the other hand, there are changes that only relate to the descriptions that can be used to truly describe a given subject. The fact that the proposition 'Socrates is admired by Alcibiades' is true at a given time but used to be false at a previous time, thus indicates a kind of change. We can certainly say that it is a change in the sense that Socrates is now the object of Alcibiades' admiration, but it is – as Geach calls it – a *mere Cambridge change*. This change does not affect Socrates *directly*: there is no *real* change about his person, only a change in a description we could give about him (and it could even be that Socrates has been dead for a long time by the time this change occurs – so that no real change of him is even possible). We could say that these are changes *for* the subject of change.[11] Every real change is thus also a Cambridge change, but not the other way around (Geach 1969, 72).

The parallel distinction I want to draw is then as follows: just as there are mere Cambridge changes (changes that are not *real* changes), we could say that there are *mere Cambridge actions* (actions that are not *real* actions). We would then have to distinguish two subclasses of actions inside a general class of Cambridge actions: the subclass of *mere* Cambridge actions, and the subclass of authentic, *real* or genuine actions. These subclasses are exhaustive and exclusive. Every genuine action (and every mere Cambridge action) is a Cambridge action, but not every Cambridge action is a genuine action. We might speak of a mere Cambridge action when no real action is in play. We are dealing with a Cambridge action involving A when we are dealing with a true proposition of the form ΦA (where Φ is an action verb or action predicate).[12] We must keep in mind that *every action* (be they *real* or *mere Cambridge actions)* can be formally described via this pattern (as when we saw earlier that every change could be seen as an alteration of the truth value of a description). That it is true that AΦs is then a necessary but not a sufficient condition for us to be dealing with a *real* action. We might then speak of a Cambridge action when we look at what an action is from a purely logical

[10] For simplicity's sake, let us say that in the following considerations, we are *not* dealing with a quantified or general proposition (which could give rise to scope ambiguities and other complications that would have to be cleared up).

[11] One must note that a *mere* Cambridge change for Socrates – in the example discussed above – corresponds to a real change on the part of Alcibiades (saying that Alcibiades has new feelings for Socrates is a *real* change on Alcibiades' part). One must always specify *who* (or what) is the subject of the change (before determining if one is dealing with a real change or with a mere Cambridge change).

[12] I suppose here that there is a categorial difference between action predicates and other predicates (like, say, quality predicates, posture predicates, passion predicates, etc.) – it is not the place here to defend this supposition, but it seems to me fairly uncontroversial to suppose that one must distinguish, when attributing something to something, different kinds of attributes, and that one of these kinds is precisely the *action* kind. As a consequence, Φ cann*ot* be '[...] is wise', '[...] is in bed', '[...] is tall' and so on. I hereby presuppose then that one could articulate one way or another what it takes for a predicate to be an action predicate.

point of view. We then have an action as soon as we can form a true sentence of the form ΦA (where Φ is an action verb or an action predicate). But the fact that something can be seen as a Cambridge action leaves open the question whether we have to deal with a *mere* Cambridge action or with a *real* action. As with the real change – mere Cambridge change distinction, an important upshot is that two (formally) logically indiscernible statements might hide an important conceptual difference: any formal description of a change is not necessarily that of a real change and any formal description of an action is not necessarily that of a real action.

That said, this distinction is not orthogonal to the distinction between one- way and two-way powers: the moon has the one-way power to create high tides but its action of creating high tides is not a mere Cambridge action, it is a *real* action of the moon (even if the moon does not have a two-way power). As I understand it, the distinction between one-way and two-way powers *presupposes* that we have *real* agents all along. By contrast, when we deal with a mere Cambridge action, the logical subject of the action is not a real agent (as when dealing with a mere Cambridge change, the logical subject of the change is not the subject of a real change).

Now, why is this distinction important? The distinction between mere Cambridge change and real change enables us to draw attention to a significant difference and allows us to specify why a mere Cambridge change is not a real change, and what might be characteristic of a mere Cambridge change. For instance, we get that a *relational* change is not a real change: if you are now on my right when you were earlier on my left, something is true of me that was not the case before, but it does not mean that I have thereby undergone a real change. To take another example, if *x* becomes a father, the description one can give of him changes while *x* does not undergo any real change (for all we know he could be ignorant of the fact). Of course, the event of *x* becoming a father *can* lead to many *real* changes in *x*'s life and in *x* himself (for instance, it can make him overwhelmed, happy, sick or stressed) so that even if becoming a father can be seen as a mere Cambridge change, it can nonetheless lead to real changes in *x* (or in other people). In the same way, *intentional* change is not a real change: if you are now jealous of me when you were not at a previous time, something is true of me that was not the case before (that I am now the object of your jealousy), but it does not mean that I myself have undergone a real change (it could very well be that I am dead when this happens). But here again, an intentional change can very well lead to a *real* change: admittedly, being loved by someone can (but does not need to) lead to substantial changes in me (in particular if it is reciprocal).

I want to suggest that similarly we could say that mere Cambridge actions have certain features. I do not mean to be exhaustive, but I would like to suggest tentatively that we are facing a mere Cambridge action in particular when someone or something is being *used* or *employed* to Φ. If I use a stone as a paperweight, one can say that the stone *prevents my documents from slipping away*. Is it a *real* action of the stone? The answer is 'no': *I* use the stone to prevent my documents from slipping away. It is something that *I* can do (and it is not something the stone can do); and the way I do it is by using the stone. It can then be an action *for* the stone. It does not mean, of course, that there are absolutely no real actions *of* the stone: it might be a real action of the stone that it leaves a mark on the paper (if it does indeed leave a mark on the paper). But the action predicate '[...] is preventing my documents from slipping away' is best understood – if the stone is its logical subject – as a mere Cambridge action. This might have to do – as in some types of mere Cambridge changes – with

the intentionality involved. In this instance, we may well have a true proposition of the form ΦA, but if A is in fact being used or employed to Φ, then we have a mere Cambridge action (and a real action from the thing that is using or employing A).[13]

Now, to return to our initial example, what does this all mean? It means that we can certainly say truly 'The plane flies' (it would be too revisionary to say that it is false or that it is a mere figure of speech); when we do, we can recognize that 'The plane' is here the subject of an action, but the question remains open as to whether the action is to be conceived as a *mere Cambridge action* or as a *real* action. In a sense of 'flying' we *use* the plane to fly (as when a pilot might say 'I flew from London to New York') – and in this case the plane is the subject of a mere Cambridge action; but in another sense of 'flying' the plane itself *can* fly (it is something that is – if conditions are met – *possible for it*). This tentative distinction might help us to move beyond the sterile opposition between the 'guns kill people' and 'guns don't kill people' positions: guns do have the one-way power to kill people, and when we say that they kill people, we neither think that they do it on their own accord, nor that the point is thereby metonymic (we do not want to really refer to the people holding guns but we do want to highlight the dangerousness of these artefacts). It is a statement about guns. In the mere Cambridge change–real change distinction, what is at stake is the question whether the change in play is the apparent subject of change's *own* change. In the mere Cambridge action–real action distinction, what is at stake is the question whether the action in play is the apparent subject of action's *own* action. When we say that hydrochloric acid dissolves bronze, it is the acid's *own* action, even if it is not a two-way power of such an acid. If you threw hydrochloric acid at my bike and pretended that you did not ruin it since the hydrochloric acid did, I might retort that in this instance, it is equally true that it did and true that this is a mere Cambridge action on the part of the acid (and a real action on your part).

For now, the important thing is that whether 'the plane flies' is a mere Cambridge action or not, it does not mean that there is any form of indirect identification (in particular a metonymy) at play in such a statement (just as there is no indirect identification when we are dealing with a mere Cambridge change). We can then conclude that an attribution of physical or natural action to a machine is not – by default – metonymic. There is no question of finding *another agent* or another subject of attribution of the action at hand. To say of the plane that it flies, of the blender that it grinds or of the fan that it blows air does not therefore encapsulate a metonymic dimension, even though it is certainly useful to distinguish one-way powers from two- way powers and to note that machines do *not have two-way powers*: the blender does not grind of its own accord; it is not up to the plane to fly or to the fan to blow air. Furthermore, it is also useful to recognize that indeed machines are agents, but can sometimes be accessories and therefore be the subject of mere Cambridge actions.

Throughout the last two sections, I aimed to show that the attribution of non-intellectual capacities to machines takes the form of a literal attribution. By introducing the two distinctions discussed in this section (one-way powers vs. two-way powers and mere Cambridge actions vs. real actions), I

[13] This distinction deserves to be explored in greater depth and will be the subject of a future study; it must be acknowledged that things become more complex when we add cases where the thing used is an agent itself capable of intentional actions.

wanted to draw attention to the fact that these distinctions gave us some leeway in such a literal attribution. A machine may well be the literal subject in the attribution of an action without being *ipso facto* endowed with a two-way power. A machine may well be the literal subject in the attribution of an action, but this action may well be a *mere* Cambridge action. Keeping this in mind, let me now turn to the question of the attribution of intellectual capacities to machines.

## The Machine as a Subject of Attribution of Intellectual Actions: The Metonymical Trap

These initial results can serve as a starting point for some reflections on the more controversial issue of whether an intellectual action can be attributed in a non-metonymic way to a machine. Can a calculator, for example, be said to *calculate* (is this really one of its actions?), can a computer program be said to *recognize*, an automated translator to *translate*? Only rarely do these questions arise in such a head-on way. The dialectic at stake more often passes through questions such as: calculators or computers undoubtedly calculate, but do they *think*? For instance, Larry Hauser (1993) argues that calculating is very much a way of thinking, that my calculator calculates *in the same sense as we do* and that as such, my calculator does indeed think.[14] The question I would like to address is thus not even raised. I would like to challenge the relevance of these attributions. The question of whether a calculator calculates is thus prior to all the skepticism that would invoke the usual solipsistic arguments (such as: we do not know what it is like for anyone or anything else to think, so we have no epistemological grounds for taking this characteristic away from the machine if it turns out that it is visibly doing the same thing as we are). This kind of skepticism was famously used by Alan Turing in his 1950 article and is nowadays prevalent (Turing 1950). Obviously, if it turns out that it is inappropriate to say that a calculator calculates, the soundness of Hauser's syllogism (P1: Calculating is thinking, P2: My calculator calculates and C: My calculator thinks) is undermined (via the rejection of P2). Wittgenstein, as we will see, has given us reasons to reject P2. Before anything, let us look at a first line of argument for doubting the capacity of a machine to calculate (or to perform any intellectual action whatsoever).

### *Classes of functional equivalence*

In order to do this, I will use Vincent Descombes' (2001) concept of 'classes of functional equivalence' or 'functionally equivalent classes'. There is a functional equivalence between two agents, Descombes tells us, when one can be substituted for the other in the context of the performance of a specific action, without affecting the actual performance of that action. A dishwasher is thus a functional equivalent of a person washing the dishes insofar as the task of washing the dishes can be performed by either one. For this functional equivalence relationship to exist between two agents, the action per- formed must be one that can be described at a certain level of generality. Furthermore, there are specific actions that only I can do, such as signing a contract (these actions cannot give rise to this class equivalence relation- ship). No one can – by

[14] This view can be seen as part of a tradition that goes back at least as far as the writings of Leibniz and his dream of a *characteristica universalis*.

definition – sign a document on my behalf (unless – of course – I have authorized it somehow, by personally signing another document, etc.). From this perspective, it is quite conceivable that some machine could perform the *exact same motions* as I do when signing a document, but this would not entail that the machine has actually signed any document (and it would certainly not entail that it has signed a document *for me*). This allows us to draw a contrast with the dishwasher: for two agents – especially if one of the agents is a machine – to be in a relationship of functional equivalence, the action at play must be described at a certain level of generality and the action must also be decomposable so that each of its components can be described outside any historical, normative or social context. The action of washing the dishes is one such action: we can describe, outside any normative or social context, what constitutes the completion of the action. Here, the action is accomplished when, say, the dishes are changed from dirty (stained) to clean (unstained). Any agent that carries out this transformation will thus have washed the dishes (from this angle, a man, a machine and a rain shower can all wash the dishes).[15] Conversely, as we have just discussed, even though we can describe the mechanical or physical decomposition of the movements involved in the signing of a document with as much precision as we wish, this alone is not sufficient to genuinely sign a document. Nobody other than the individual who is supposed to sign would be able to sign. In order to sign a document, it is, therefore, necessary that it is the person who is supposed to sign who does so, and it is, therefore, impossible for anyone or anything else to do so (not physically but logically impossible).

A reason why it is logically impossible for someone other than me to sign a document is that the action of signing a document is a normative action. It therefore relates to the customs and conventions of a given society. These normative or conventional aspects are crucial in determining the nature of intellectual or intentional actions. They are crucial in that they allow us to fully grasp why such actions are not amenable to a relation of functional equivalence, and therefore to understand why they cannot literally be performed by machines. Let us now say a little bit more about this normative, social and historical dimension of intellectual actions.

## *The historical and social dimension of intellectual actions*

On the one hand, we have actions that are insensitive to the historical or social context for their attribution, while on the other hand, we have actions that are essentially dependent on this historical and social context. It is one of Wittgenstein's major insights to have drawn attention to this contrast. One of Wittgenstein's philosophical breakthroughs consists in understanding that calculation (and more generally every other intellectual or mental action) is precisely the sort of action that *cannot be described in isolation from its normative and social framework*. If this is so, then it is *impossible* for intellectual actions to constitute the kind of actions that can give rise to these functional equivalence relations. Hence, it is only if an action meets certain criteria that it can, for instance be described as one of 'calculating' – criteria that, in fact, go beyond the mere physical framework (and relate crucially to the institutions and norms of calculation). In the *Philosophical Investigations* (Wittgenstein 2009: PI §87), Wittgenstein thus comments that 'To follow a rule, to

[15] Any discomfort there may be at this point is alleviated by the considerations raised in the previous section (the contrast between one-way and two-way powers and the distinction between real actions and mere Cambridge actions).

make a report, to give an order, to play a game of chess, are *customs* (usages, institutions)'. All these activities are indeed normative activities in the sense that they can only take place within a society (which has its very own norms and practices). As a result, these actions can only be carried out by individuals who are familiar with the existence of these norms and practices and who are therefore in a position to make them their own (by following them or, on the contrary, by disregarding them or by failing to follow them). Calculating thus requires that there already exists a practice of calculation. A related matter – developed on multiple occasions in Wittgenstein's work – is that the practice of calculating (of signing, buying, etc.) itself hinges on a *consensus* (Wittgenstein 1967: RFM, 94), that is, on the recognition by the members of a society of its function and role. Thus, in order to grasp and fully account for an intellectual action, it seems necessary to refer to the *context* of the action, that is, to bring in external factors to the mere gestures or mechanical movements that may otherwise constitute it, just like what we said earlier about the gestures that are made when signing: these gestures, admittedly, have their importance but are by no means sufficient to constitute a signature.

The social, historical and normative dimension of intellectual actions thus explains why these actions cannot be reduced to mere physical movements – and why they cannot, therefore, be amenable to a relation of functional equivalence. There can indeed be no functional equivalent for an individual's action of signing a document or performing a calculation, whereas there can, on the contrary be such an equivalent for an individual's action of washing the dishes. Thus, the calculator does not calculate (whereas the dishwasher does wash the dishes) and is just a way among others *for human beings* to calculate – and so, to exercise their intelligence. It is now this idea of acting intelligently *with* a machine that requires us to say a few words since this aspect of our way of performing intellectual actions might sometimes seem to blur the distinction between the thing that follows a rule and the thing used to follow a rule.

## *Acting intelligently with a machine*

One of the obstacles to thinking about machines as mere metonymical subjects of intellectual actions is that we can certainly perform some normative, social or intellectual actions *with* them (and that they, therefore, help us to perform these tasks). No one today is surprised by the idea of an 'electronic signature', but an electronic signature is not about a machine signing *for me*, it is about me signing *with* the machine (rather than with a pen). When I electronically sign a document, the historical and social criteria are fully at play (remarkably if *someone other than me has electronically signed it*, there has been no signature *at all*). To take another example: there would be something fantastic about developing a machine for selling. 'Selling' is not one of those actions that can be described outside of any institutional context: selling presupposes, amongst other things, the institution of private property, money and so on. I may indeed sell something with the help of a machine, but it would make no sense to say of a machine that it has sold anything. 'Selling' cannot therefore give rise to relations of functional equivalence as we defined them earlier. That we can sell or sign *with* a machine does not imply that a machine can sell or sign. The same applies to intellectual actions: that we can calculate with a machine does not imply that a machine can calculate. The subject of an intellectual action when a machine is involved can therefore literally only be the individual using the machine – not the machine itself. This brings us back to our metonymical trap. The temptation

is great, in the case of intellectual actions, to consider that the machine *could* be the subject of such actions, but it is a temptation that must be resisted since no clarity could come out of it: thinking that it could sell, promise or sign anything is as absurd as thinking that a machine could believe, hope, dream and so on. How could, for example, a *promising machine* be remotely intelligible?

## Conclusion

There is a shortcut in saying that a machine calculates that may lead us astray. We can certainly calculate *with* a machine, but there is no sense in saying that a machine itself calculates (any more than there is any sense in saying that it signs, sells, etc.). This position has its roots in the writings of Wittgenstein and is often overlooked on the philosophical battlefield.

One can assign certain types of action to machines. Some actions can be performed indistinctly by a human and a non-human agent. An intellectual action, on the other hand, can only be carried out by an agent with a human life. If, during a hailstorm, my calculator displays '8' after being hit successively by small pellets of ice on the '5', '+', '3' and '=' keys, it would be incongruous to say that my calculator has calculated the result of this sum (as it would also be incongruous to say that the hailstorm did). Wittgenstein takes a similar example (Wittgenstein 1967: RFM, 133) while attempting to show that under such circumstances one could not reasonably consider that someone or something actually carried out a calculation. Indeed, none of the things involved (the hailstorm, the pellets of ice) has any awareness of the institution and norms of calculation. On the other hand, it is not incongruous to say that *I calculate* this sum when I press these keys myself (even if I calculate *with a calculator*). The metonymical trap then consists in taking the tool used to perform the action for the literal subject of the action.

## References


Alvarez, M. 2013. 'Agency and Two-Way Powers'. *Proceedings of the Aristotelian Society* 113: 101–121.

Aristotle. 2016. *Metaphysics*. Translated by Reeve, C. D. C. Indianapolis, IN: Hackett Publishing Company.

Bennett, M. R. and Hacker, P. M. S. 2023. *Philosophical Foundations of Neuroscience,* 2nd edn. New-Jersey: Wiley Blackwell.

Descombes, V. 2001. *The Mind's Provisions: A Critique of Cognitivism. Translated by Schwartz, S. A. Princeton*: Princeton University Press.

Geach, P. T. 1969. *God and the Soul*. London: Routledge.

Hacker, P. M. S. 2010. *The Categorial Framework*. Oxford: Wiley Blackwell.

Hacker, P. M. S. 2013. *The Intellectual Powers*, Oxford: Wiley Blackwell.

Hauser, L. 1993. 'Why Isn't My Pocket Calculator a Thinking Thing?'. *Minds and Machines* 3(1): 3–10.

Lakoff, G. and Johnson, M. 2003. *Metaphors We Live By*, 2nd edn. Chicago: The University of Chicago Press.

Littlemore, J. 2015. *Metonymy: Hidden Shortcuts in Language, Thought and Communication*. Cambridge:

Cambridge University Press.
Thomasson, A. L. 1998. *Fiction and Metaphysics*. New York: Cambridge University Press. Turing, A. M. 1950. 'Computing Machinery and Intelligence'. *Mind LIX* 236: 433–460. Walton, D. 2008. *Informal Logic, a Pragmatic Approach*, 2nd edn. New York: Cambridge University Press.
Wittgenstein, L. 1958. *The Blue and Brown Books*. Oxford: Basil Blackwell.
Wittgenstein, L. 1967. *Remarks on the Foundations of Mathematics*. Cambridge: The MIT Press.
Wittgenstein, L. 1980. *Remarks on the Philosophy of Psychology*, vol. II. Oxford: Basil Blackwell.
Wittgenstein, L. 2009. *Philosophical Investigations*, 4th edn. Oxford: Wiley Blackwell.