The Prismatic Shape of Trust

1. A Theoretical Approach

Il prisma della fiducia

1. Approcci teorici

TEORIA

Rivista di filosofia
fondata da Vittorio Sainati
XXXIX/2019/1 (Terza serie XIV/1)

Edizioni ETS

«Teoria» è indicizzata ISI Arts&Humanities Citation Index e SCOPUS, e ha ottenuto la classificazione "A" ANVUR per i settori 11/C1-C2-C3-C4-C5.

La versione elettronica di questo numero è disponibile sul sito: www.rivistateoria.eu

Direzione e Redazione: Dipartimento di civiltà e forme del sapere dell'Università di Pisa, via P. Paoli 15, 56126 Pisa, tel. (050) 2215500 - www.cfs.unipi.it

Direttore: Adriano Fabris

Amministrazione: Edizioni ETS, piazza Carrara 16-19, 56126 Pisa, www.edizioniets.com, info@edizioniets.com - tel. (050) 29544-503868

Abbonamento: Italia € 40,00 (Iva inclusa); estero € 50,00 (Iva e spese di spedizione incluse) da versare sul c.c.p. 14721567 intestato alle Edizioni ETS.
Prezzo di un fascicolo: € 20,00, Iva inclusa.
Prezzo di un fascicolo arretrato: € 30,00, Iva inclusa.

L'indice dei fascicoli di «Teoria» può essere consultato all'indirizzo: www.rivistateoria.eu. Qui è possibile acquistare un singolo articolo o l'intero numero in formato PDF, e anche l'intero numero in versione cartacea.

Iscritto al Reg. della stampa presso la Canc. del Trib. di Pisa n° 10/81 del 23.5.1981. Direttore Responsabile: Adriano Fabris. Semestrale. Contiene meno del 70% di pubblicità.

I numeri della rivista sono monografici. Gli scritti proposti per la pubblicazione sono double blind peer reviewed.
I testi devono essere conformi alle norme editoriali indicate nel sito.

TEORIA

T

Rivista di filosofia
fondata da Vittorio Sainati
XXXIX/2019/1 (Terza serie XIV/1)

The Prismatic Shape of Trust

1. A Theoretical Approach

Il prisma della fiducia

1. Approcci teorici

Edizioni ETS

Contents / Indice

II. *Philosophy, Knowledge, and the Sciences*
II. *Filosofia, conoscenza e riflessione scientifica*

a cura di Giovanni Scarafile

T

Premise / Premessa

This issue of «Teoria» has been divided into two sections. The first includes some of the contributions that were selected in response to the call for papers on the subject of *The Concept of Trust*. These are international contributions that address the issue of "trust" from a theoretical perspective, taking into account the many areas in which this specific aspect of inter-human relations plays a role. Some of the authors introduced here then went on to discuss their theses during a study day, titled *Il prisma della fiducia (The prism of trust)*, which took place at the University of Pisa on 11th December 2018.

In the second section of this issue we have published some papers that were presented at the International Convention of *Philosophy, Knowledge, and the Sciences*, held at the UNISER, University of Pistoia, on 4th and 5th June 2018. The convention – organised by the Doctoral School of Philosophy for the Universities of Florence and Pisa, as well as by UNISER – saw the participation of the following speakers: John Schellenberg (Mount Saint Vincent University, Canada), Paolo Crivelli (Université de Genève), Jean-Michel Salanskis (Université Paris Nanterre), Christian Wüthrich (Université de Genève), Marcel Weber (Université de Genève), Carole Talon-Hugon (Université Nice Sophia Antipolis), Petar Bojanić (Institute for Philosophy and Social Theory, University of Belgrade), Renaud Barbaras (Université Paris 1 Panthéon-Sorbonne), as well as professors and students of the Doctoral School in the role of *discussants*.

The result, in both these sections, is a series of essays that have not only been evaluated and discussed for publication purposes through a *double blind peer review* process, but which, moreover, are the result of comparison and common debate developed between scholars, both young and old,

 DOI: 10.4454/teoria.v39i1.53

who are specialised in this field. A similar format will also be adopted for the next issue of «Teoria», which will feature the second part of the papers selected in relation to the theme of trust. These will deal with the question from a more specifically historical point of view, analysing some important authors on this topic.

Questo fascicolo di «Teoria» si compone di due sezioni. La prima raccoglie una parte dei contributi selezionati in risposta al call for papers sul tema *The Concept of Trust*. Si tratta di contributi internazionali che affrontano il tema della "fiducia" da una prospettiva teorica e tenendo conto dei molti ambiti in cui questa specifica qualità delle relazioni interumane viene a giocare il proprio ruolo. Alcuni degli autori che qui presentiamo, poi, hanno discusso le loro tesi in una giornata di studi, dal titolo *Il prisma della fiducia*, che si è svolta all'Università di Pisa l'11 dicembre 2018.

Nella seconda sezione del fascicolo sono invece pubblicati alcuni testi presentati al Convegno internazionale su *Philosophy, Knowledge, and the Sciences*, svoltosi all'UNISER, Polo Universitario di Pistoia, il 4 e 5 giugno 2018. Il convegno – organizzato dalla Scuola dottorale in Filosofia delle Università di Firenze e di Pisa, e dall'UNISER – ha visto la partecipazione, come relatori, di John Schellenberg (Mount Saint Vincent University, Canada), Paolo Crivelli (Université de Genève), Jean-Michel Salanskis (Université Paris Nanterre), Christian Wüthrich (Université de Genève), Marcel Weber (Université de Genève), Carole Talon-Hugon (Université Nice Sophia Antipolis), Petar Bojanić (Institute for Philosophy and Social Theory, University of Belgrade), Renaud Barbaras (Université Paris 1 Panthéon-Sorbonne) e, in qualità di *discussant*, dei docenti e degli studenti della Scuola dottorale.

Il risultato di entrambe le sezioni è costituito da una serie di saggi che non solo sono stati valutati e discussi ai fini della pubblicazione attraverso un processo di *double blind peer review*, ma che, anche e soprattutto, sono il risultato di un confronto e di un dibattito comune sviluppatosi fra studiosi vecchi e giovani competenti in materia. Un analogo percorso verrà seguito anche per il prossimo fascicolo di «Teoria», che conterrà la seconda parte dei testi selezionati in relazione al tema della fiducia. Essi affronteranno tale questione da un punto di vista più propriamente storico, analizzando alcuni autori importanti per la sua trattazione.

Adriano Fabris, Giovanni Scarafile

I.
The Prismatic Shape of Trust
1. *A Theoretical Approach*

I.
Il prisma della fiducia
1. *Approcci teorici*

a cura di
Adriano Fabris

T

Scientific Experts and Citizens' Trust: Where the Third Wave of Social Studies of Science Goes Wrong

Pierluigi Barrotta, Roberto Gronda

According to a familiar approach, in cases of technological decision-making – i.e., in those cases in which the subject-matter of political deliberation presents a scientific element as one of its essential features – a clear-cut distinction should be drawn between the technological and scientific moment, on the one hand, and the socio-political one, on the other. First of all, the former is taken to be conceptually prior to the latter. Secondly, the technological and scientific moment is understood as being up to the experts: their aim is said to be the quest for facts, the search for an explanation, the construction of reliable technological tools, and so on. After that, once reliable technological tools have been produced by the experts, the socio-political moment takes the stage for the purpose of setting policies to implement: society, through its representatives, asks the experts about the means to reach the ends and values established by the society itself. The clear-cut distinction between techno-scientific and socio-political moment therefore goes hand in hand with the distinction between facts and values. More precisely, it is the epistemological soundness of the fact-value dichotomy which grounds the idea that in any well-conducted policy the two moments should be kept separated.

Such conception has been strongly criticized by the proponents of Social Studies of Science (henceforth, STS). They argue that the alleged objectivity of scientific experts is a myth, and, consequently, that there is no sound epistemic reason to trust them. In reality, the facts on which their knowledge is ultimately based are constructed by society. The task of sociology of knowledge is precisely that of deconstructing those facts, for the purpose of showing the role played by social values and interests in the process of their constitution.

 DOI: 10.4454/teoria.v39i1.54

The growing de-legitimation of scientific experts has reached such a point that some sociologists of science – born and raised within the STS paradigm of research – have realized that that tendency should be countered. This is the aim of the Third Wave of Social Studies of Science, launched by Harry Collins and Robert Evans. The Third Wave, which Collins and Evans champion, aims at defending all the values, truth included, which define science as a specific "life-form". As they remark in their recent book *Why Democracy Needs Science*: «We desperately need to preserve the moral imperative that guided science under Wave One», that is, under the traditional image of science as a distinctively epistemic enterprise (Collins and Evans 2017: 77). They argue that, even though it should be acknowledged that the claims of scientific experts to provide society with reliable knowledge have proven unwarranted, it is nonetheless possible to defend the importance of their role in society on moral grounds.

The goal of the present essay is to analyze and criticize Collins and Evans's view. We argue that, no matter how ingenuous it might be, their proposal is highly debatable. Their argument relies on the assumption – which we firmly reject – that the objectivity of scientific knowledge presupposes a foundationalist epistemology, as a consequence of which an epistemic account of expertise cannot be advanced. We disagree on this point, and we also believe that a defense of the notion of expertise on purely moral grounds is both descriptively and normatively unsuccessful, and should therefore be rejected.

We start off with a conceptual consideration. Collins and Evans think that the notion of expertise should be viewed as substantial. They argue that being an expert is a property that a person possesses independently of the fact that he or she is acknowledged as an expert. On the contrary, we adopt a relational account, centered on the notion of trust, which paves the way for the distinction between scientists and experts. We agree with Collins and Evans that being a scientist – being a scientist by *profession* – should be treated as a substantial notion. However, we maintain that the *status* of expert necessarily implies a relation to a group of persons who choose to trust that particular scientist as a reliable source of knowledge. A scientist becomes an expert when she is trusted by a group of laypeople.

Such relational account, inspired by pragmatism, enables us to vindicate the intrinsically epistemic character of the notion of expertise. Laypeople do not consult experts because they are willing to preserve the moral values embodied in the latter's form of life, but rather because they have some reason to trust them as reliable source of knowledge in light of

the particular problems at stake in technological deliberation. The notion of trust so conceived cannot be defended on purely moral grounds, but has to be accounted for in strictly epistemic terms. Our approach, centered on the Deweyan idea of problem-solving, aims to show that it is possible to defend an epistemological conception of expertise without relapsing into the traditional – and rightly criticized – dichotomy between techno-scientific and socio-political moments.

In the first section of this article, we lay out Collins and Evans's Three Waves of STS. In the second section, we discuss and analyze their moral defense of science and expertise. In the third section, we criticize Collins and Evans's view, and we challenge the tenability of their position. Finally, in the fourth section, we present our pragmatist account of expertise, and we argue that the notion of expertise cannot be understood apart from the notion of trust.

1. *Three Waves of Social Studies of Science*

In numerous articles and books, Collins and Evans have suggested dividing the history of STS in three great moments or waves. Among other things, those three phases represent three different ways in which the relationship between science – and expertise – and democracy can be framed.

The First Wave, which corresponds to the dominant paradigm prior to the 60s, was characterized by the belief that it was possible to provide an epistemological justification of scientific inquiry. The First Wave believed it was possible to single out clear-cut criteria of demarcation separating science both from non-science and from other human activities such as politics or propaganda. Science was characterized as that enterprise exclusively concerned with the discovery of truth, and it was understood as free from moral and social values. Consequently, the theories advanced by scientific communities were taken to be genuine instances of knowledge. On this basis, Wave One was able to offer a simple and straightforward account of the relation between science and society. Since scientists are exclusively concerned with the discovery of truth, their activity is not burdened with social biases and moral prejudices. The fact that scientists are value-free entails, therefore, their reliability as experts. The judgments of scientists are exclusively responsive to how things are in themselves, as a consequence of which citizens are justified in trusting scientific experts. Even more radically, on these bases there is no sound reason why citizens

should not trust them. Trust from citizens is not something that scientists should earn; it is a by-product of the methodological assumptions of their disciplines.

Wave One's conception of science is no longer believed: it strikes us as naïve and over-simplistic. Starting from the seminal book of Thomas Kuhn, *The Structure of Scientific Revolutions*, both philosophers and sociologists of science have turned their attention to the non-epistemic factors that make scientific knowledge possible. The distinctive feature of Wave Two is the harsh criticism of science's demand for objectivity, and its motto is «distance leads to enchantment». When the curtain is raised, and the actual behavior of scientists is empirically investigated, it is easy to see, so the argument goes, that knowledge is less the result of a confirmation of theories by evidence than the outcome of rhetorical strategies of persuasion. Discovery of truth is therefore nothing but a misleading name for the social process of negotiation of what counts as valid within a specific community. As Collins and Evans remark: «under Wave Two, science is eroded as non-scientific values encourage new kinds of behavior. […] The view associated with Wave Two is that the truth of the matter cannot be found, that there are only interpretations and perspectives» (Collins and Evans 2017: 108 and 40).

In this scenario, there is no reason why citizens should trust scientific experts. Indeed, what goes under the name of scientific knowledge is made of the same stuff as political deliberation. Science is loaded with moral and social values; consequently, scientific experts are in no better position to tell citizens what ought to be done. Science and politics are negotiation and compromise through and through: «science is politics pursued by other means» (Latour 1983: 168).

The Third Wave of STS launched by Collins and Evans aims to counteract and defuse the most radical conclusions reached by Wave Two. In particular, it aims at defending the role of scientific experts in democracy (whence the title of their book, *Why Democracy Needs Science*) without relapsing into the naïve image of science formulated by Wave One. The most interesting aspect of Collins and Evans's proposal is that they believe that the distinctive values of science can be preserved within the Wave Two approach. They formulate this insight by saying that while Wave Two «showed that Wave One was intellectually bankrupt», Wave Three should be seen as a development and refinement of Way Two rather than as an attempt to reject its premises (Collins and Evans 2002: 240).

Wave Three agrees therefore with Wave Two on almost everything the

latter has said about the nature of scientific knowledge (Collins and Evans 2017: 11). They both hold that science is not value-free, that scientific knowledge should properly be seen as the result of a process of social negotiation, and that scientists do not have any privileged access to reality. The only point of divergence between the two concerns their normative position. While Wave Two leans towards more democratization, Wave Three purports to reintroduce a set of distinctions, on whose basis «to preserve the idea of expertise as specialist knowledge and to find a better way of analysing and managing the trade-offs between expert authority and democratic accountability» (Collins and Evans 2017: 11).

So, they remark, «Wave Three involves finding a special rationale for science and technology even while we accept the findings of Wave Two» (Collins and Evans 2002: 44; quoted in Collins and Evans 2017: 100). In order to do so, Collins and Evans distinguish between two different problems, that of legitimacy and that of extension. Wave Two was mainly concerned with the problem of legitimacy: its goal was to show that, once it is acknowledged that «the apparently neutral and objective advice provided by technical experts cannot have the unquestionable epistemological authority it claims», a more reliable procedure can be achieved if a «wider range of perspectives and experiences» is allowed to be represented into the decision-making process (Collins and Evans 2017: 13).

On the contrary, Wave Three is concerned with the problem of extension, which was left unanswered by Wave Two. Indeed, the latter has merely shown that more "subjects" than the experts are legitimated to participate in the decision-making process; it has not addressed the issue of the scope and limits of participation. To properly answer this question, a normative stance has to be adopted, which provides criteria for inclusion and exclusion, and, in doing so, also settles once for all the problem of legitimacy. As Collins and Evans remark:

> [T]he solution to the problem of legitimacy is also the solution to the problem of extension: all the "right" people will have a say in the technical debate, and those who have no relevant specialist expertise will contribute as citizens participating in existing democratic institutions without pretending to be, or being described as, experts (Collins and Evans 2017: 14).

However, in order not to betray the spirit of Wave Two, those criteria cannot be epistemic. A different route must be taken.

2. *Collins and Evans's Moral Defense of Expertise*

It has been said that the Third Wave of STS aims to preserve the idea of expertise as specialist knowledge, and that the argument in support of this view cannot be epistemic. Wave Two, Collins and Evans write, has shown that there is nothing special about science: as a consequence of that, they notice, it is now very difficult to defend science «on the grounds of its truth and utility» (Collins and Evans 2017: 19). However, it is possible to take a moral road, and defend science on the grounds of its contribution to the values of a community. The key point here is to acknowledge that, even though it is true that science cannot reach truth, the values that it embodies and exemplifies are nonetheless eternal.

It is not easy to find in Collins and Evans's work an explicit formulation of the line of thought that is supposed to warrant that thesis. In some passages, they seem to derive it directly from the fallibilistic view of science. In particular, they seem to maintain that according to fallibilism – which is commonly held as the standard position in philosophy of science – no foundation of our best scientific knowledge can be provided, which entails that there is no sound epistemic reason to trust science. Put in this way, the argument is untenable: the rejection of foundationalism and the consequent adoption of a fallibilistic perspective do not amount to discharging any possible form of objectivity. Epistemologically speaking, this is a *non sequitur*. We will therefore try to outline a plausible argument that, we believe, could be accepted by Collins and Evans as faithful to their intentions. Only after having clarified the argument, we will go on to criticize their position.

As said, Collins and Evans's starting point is the thesis of the fallibilistic nature of science. We have also remarked that that thesis is not strong enough to directly support the conclusion that they would like to draw from it. But let's put the matter in another, slightly different way. First of all, assume the validity of the pessimistic meta-induction. According to this view, since all the scientific knowledge that was taken as true in the past has been later shown to be false, we should have the humility to admit that our best scientific theories will very likely turn out to be false in the future. Indeed, there is no evidence that current scientific theories are substantially different from the ones believed in the past; so, they may well share the same fate. At the end of the day, Newtonian mechanics seemed correct for so long, and yet has now been shown to be false and has been replaced by Einstein's theory of relativity.

Pessimistic meta-induction is a highly questioned concept, and is far

from being uncontroversial. Nonetheless, it is not wholly implausible, and can be argued for with some success. Note that, in order for the argument to be consistent, it is necessary to take pessimistic meta-induction in its more radical form, as excluding truth-approximation. Indeed, if contemporary scientific theories turn out to be less false than their predecessors, it would be still possible to introduce an epistemic element in the context of evaluation: the epistemic value of avoiding error is almost as important as that of reaching the truth. For the sake of this argument, we will also assume this radical version of pessimistic meta-induction as plausible, even though we are very dubious about its soundness.

The second assumption of this argument elaborates on the first, and can be formulated as follows. We know that past scientific knowledge turned out to be false, and we also know that current scientific knowledge will turn out equally false; nonetheless, we still hold science dear, and we are ready to defend it from the vicious attacks of its opponents. So, for instance, we are willing to defend evolutionary theory against the claims of Creationism, even though it is very likely that both are false. It follows therefore that the reasons why we are led to defend and safeguard science are not epistemic, since we are committed to its preservation within our society independently from its truth.

This second assumption seems more plausible – at least *prima facie* – than the first one since its content has a strong and unquestionable factual component. It reports that a significant number of citizens in Western societies are ready to defend science against those who are willing to deny its importance for our form of life. This is a sociological – i.e., empirical – fact, and, consequently, we as philosophers take it for granted. Collins and Evans might be ready to say that if the plausibility of the first assumption is admitted, the truth of the second one is hardly questionable, since the latter can be seen as a corollary of the former. We do not agree with them, but we will postpone the examination of this issue until the argument has been settled.

So, it may be asked, if this account is correct, what is the rationale behind our choice in favor of science? The last step of the argument is the core of Collins and Evans's theoretical proposal – which they label "elective modernism". They maintain that since we are interested in preserving science even in the absence of sound epistemic reasons, the motives of our decision should be of a different kind, namely, of a moral kind. More precisely, the reason why we are interested in having science in our society is that the life-form of science embodies values which we want to preserve. According to Collins and Evans, contemporary Western societies are

undergoing a progressive erosion of their distinctive values, in great part as a consequence of free-market capitalism. In this scenario, they argue, science is one of the few remaining fortresses of morality.

Science ceased therefore to be conceived of as the privileged source of knowledge, and become the «fountainhead of values» (Collins and Evans 2017: 19). But what are these values that are deemed as worthy of preservation? Collins and Evans are explicit that they do not have a new list of values to propose; they rely on Merton's classical analysis, and translate the latter in terms of the notion of "formative aspirations of science". "Formative aspiration of science" is a heuristic tool which refers to the set of normative constraints that should be satisfied in order for an action to count as a *scientific* action, and for an individual to count as a *scientist.* The notion of aspiration highlights the fact that an individual need not be successful in satisfying those constraints; indeed, this would be a too restrictive condition. It is enough that her actions are guided by the values of science: these are *observation*, *corroboration*, *falsification*, and the Mertonian norms of *communism, universalism, disinterestedness*, and *organized skepticism.*

Because no epistemic justification of those norms is believed to be possible, Collins and Evans do not attempt to demonstrate their validity. They are content to appeal to the moral conscience of the citizens of Western society, asking them whether they prefer to live in a society in which expertise is respected and defended, and information is shared, discussed, weighted and criticized; or rather in a society in which no distinction is drawn between experts and lay people, and information is kept in the hands of the few.

It is clear, and it is difficult to disagree with Collins and Evans on this point, that we prefer to live in the first type of society. Nonetheless, if elective modernism is true, our preference turns out not to be grounded on proofs and demonstrations: epistemic justification makes way for moral persuasion.

3. *A Criticism of Collins and Evans's Moral Strategy*

Up to now, we have limited ourselves to reconstructing Collins and Evans's argument. It is about time to assess its validity. In this section, we will list some objections to their proposal, for the purpose of showing why we take it to be seriously flawed. These objections will then pave the way for the formulation of our pragmatist account of expertise, which will be outlined in the next chapter.

The first objection is epistemological, and is concerned with Collins and Evans's criticism of foundationalism. Collins and Evans seem to hold that any possible form of objectivity – no matter how it can be conceived of – is essentially interwoven with the foundationalist project, to the effect that in order for knowledge to be true, it must be grounded on some indubitable set of principles or data. We do not have enough space here to delve into a detailed analysis of this issue, so we limit ourselves to a sort of sociological remark. Foundationalism is now hardly a mainstream position in the contemporary philosophical landscape, but, nonetheless, attempts to come up with a consistent theory of objectivity are a daily occurrence. This is due to the fact that foundationalism and objectivity of knowledge are distinct concepts: at best, foundationalism is one of the manifold ways in which the objectivity of scientific knowledge can be accounted for. Things are much more complicated than Collins and Evans think they are.

The second objection is directed against the conclusion of the argument, i.e., the idea that science can be defended on purely moral grounds. Two points are at stake here. First of all, Collins and Evans seem to commit a fallacy of abstraction. In general terms, the latter takes place when a certain complex phenomenon is investigated and analyzed from a specific point of view, and the results of the investigation are identified with the whole phenomenon. In the case under discussion, it is evident that science can be investigated as a form of life, and it is also evident that it is possible to single out some values as distinctive of scientific activity. There is nothing wrong in treating the values of science as formative aspirations of its practitioners; similarly, it is completely legitimate to defend those values on moral grounds. Nonetheless, that does not mean that the moral dimension can be severed from the epistemic one, and taken as autonomous.

Secondly, Collins and Evans are not content to sever the moral aspect of science from its epistemic dimension. They also place the two aspects in contrast with each other[1]. In doing so, they consciously refuse to employ epistemic resources to strengthen their argument. It goes without saying

[1] For the sake of fairness, it should be noted that Collins and Evans do not rule out the possibility of an epistemic defense of science. They are clearly not committed to it, but they also argue that their elective modernism makes room for this kind of approach. They suggest to read their proposal as adding a second arrow to the quiver of those who are interested in defending the role and function of science in contemporary democracies. It is nonetheless very difficult to see how this is actually possible, since Collins and Evans are explicit in rejecting the validity of the epistemological analysis of science. In our opinion, their concession sounds less like a genuine theoretical option than a rhetorical *captatio benevolentiae*.

that theirs is a bold choice; however, its consequences are puzzling. Here is what Collins and Evans write about their moral defense of the value of observation: when it is said that those who have observed something in a systematic way are «[a] *better* source of opinion that those who have not», they remark, the italicized better «cannot mean "more efficacious"» since «if it did we would have a foundational justification»; consequently, they conclude, «[better] does not mean better at anything, it just means better» (Collins and Evans 2017: 20).

As is evident from this quotation, Collins and Evans argue that there is no epistemic reason why we should prefer observation over mere guessing; the only sound reason is that we should prefer to live in a society were people do observations, are skeptical about their conclusions, are open to discussion, and are willing to falsify their beliefs at certain occasions. That preference is moral; it has to do with the way in which we would like to conduct our lives. It has nothing to do with the epistemic credentials of those acts.

But is it truly so? Is the picture of science that Collins and Evans draw plausible? Part of our perplexities are related to, or depend on, the epistemological confusion that we have criticized above, so we won't repeat them once again. But there is something more to it. Let's take the idea of the fallacy of abstraction seriously. It seems clear to us that we have good reasons to prefer observation over mere guessing, and we agree with Collins and Evans that some of these reasons are moral. After all, scientific observation is grounded on the virtue of carefulness, which is a trait of a reliable and responsible character. Since observation is evidence of a good character, we prize it, and we are ready to defend that activity on moral grounds. Such entanglement of the epistemic and the moral is not problematic for our argument: we accept it unhesitatingly. The point is: are these moral reasons as autonomous from the epistemic ones as Collins and Evans would need them to be in order to justify their conclusion?

We think not. Imagine a strongly counter-factual situation in which, because of radically different laws of nature, observation did not have any epistemic value. Suppose, for instance, that the past continuously changed in ways which were unpredictable to us. Consequently, what we have observed at t cannot count as evidence at t' because things are now different from how we saw them. In this case, would observation be defended as a moral value? It seems that this can hardly be the case. From a genealogical point of view, it is very difficult to believe that mankind would have developed a genuine interest in observation if the latter had been completely ineffective. At the end of the day, if inspecting the viscera of birds

had proved itself a reliable method to forecast the future, human beings would have continued to consult the haruspices for information. The fact that observation has been preserved in the course of evolution seems therefore unaccountable unless we have recourse to epistemic values. Nonetheless, Collins and Evans are committed to what we may call the thesis of the dispensability of the epistemic. Accordingly, that move is not open to them: they are compelled to use exclusively moral resources. Honestly, we do not see any possible way out of this predicament.

However, it is fair to remember that Collins and Evans do have at least one other argument in support of their conclusion, which should – or is at least intended to – corroborate the idea of the dispensability of the epistemic. Here is their argument.

Elective modernism is concerned with technical decision-making. Consequently, it is at this level that the validity of Collins and Evans's approach should be properly assessed. Now, when the focus of analysis is shifted from scientific research to technical decision-making things change dramatically. Indeed, in the case of technical decision-making, experts are asked to answer questions that are urgent and decisive for society, without having time to do further investigations and defer their answer. It is a fact that when they have to act under these conditions, experts are often wrong: the opening pages of Collins's book *Are We All Scientific Experts Now?* provide an impressing overview of the errors of experts, from mad cow disease to 2008 financial crisis.

Technical decision-making shows, therefore, that the traditional, epistemological image of science is a myth. From a strictly epistemic point of view, scientific experts are not as trustworthy as we may want them to be since there is strong empirical evidence that they make a lot of mistakes. In addition, the consequences of such mistakes are not confined to the laboratory, but affect the lives of thousand and thousand of people.

Despite all of this, elective modernism wants to defend the positive role of science in society. However, the epistemic track record of science is not strong enough to provide a consistent argument in its support. Consequently, we had better go the moral route.

This argument is ingenious. It introduces some new concepts that actually change the agenda of discussion. In particular, the shift from scientific research to technical decision-making is theoretically fertile, and also shows a promising direction to explore. All that said, however, we still believe that Collins and Evans do not succeed in satisfactorily arguing for the validity of their elective modernism. Their argument is shaky at best.

First of all, it relies on a selective induction from negative cases only. Nobody is willing to deny that scientific experts are often wrong – even though the reasons for their mistakes should be carefully investigated. However, it is simply not true that experts are always wrong: in a technical decision-making scenario the risk of error is undoubtedly enhanced, but it is exaggerated to conclude that expert advice is epistemically unreliable. In addition, much of scientific knowledge is not deterministic: the fact that a singular case may happen to be in contrast with a set of general laws held by the scientific community does not count as evidence of the epistemic unreliability of those laws. On this point Collins and Evans are simply too rash.

Secondly, if their argument were correct, some unfortunate moral consequences would follow. Suppose Collins and Evans are right, and assume that there is no epistemic ground for scientific expertise, but only some kind of moral persuasion. The search for truth – which is the «fundamental formative aspiration of science», according to Collins and Evans – would therefore turn out to be an illusion since, as Wave Two has shown, «the truth of the matter cannot be found», and «there are only interpretations and perspectives». Note that Collins and Evans are happy with this view: it is true that their aim is to somehow counteract the «corrosive effect of Wave Two», but they do not question the validity of its conclusion (Collins and Evans 2017: 40). There is no truth of the matter.

How is it possible? Collins and Evans ask us to distinguish truth conceived of as the value which should justify scientific inquiry (let's call it, Truth with capital T) from the notion of truth as is usually employed by scientists to characterize their own particular form of life (truth with lowercase t). While Truth must be gotten rid of, the concept of truth is essential for science as a form of life. Indeed, if scientists do not believe that they are actually succeeding in discovering the truth of reality, their work as scientists would be substantially impossible. As Collins and Evans explicitly remark:

> One cannot do good science without disbelieving social constructivism. Individual scientists have to believe they are seeking the truth and that there is a chance of finding it, even while social scientists insist it is the social group that ultimately determines what counts. Furthermore, scientists must ignore the social constructivists if the formative aspirations of science on which this entire thesis turns are to be robust (Collins and Evans 2017: 76).

Collins and Evans do not see any trouble with this sort of self-deception.

On the contrary, they argue that it is necessary in order to preserve the source of values that is science. This assumption is highly disputable, but let's accept it for the sake of discussion[2]. At the end of the day, one may even argue that such self-deception is for the greater good, since scientists are thus given the chance to live a valuable life in a privileged environment. So, let's concede that this deception is benign.

But consider another kind of deception: imagine a society in which experts – who are not a reliable epistemic source of information – are nevertheless still consulted by citizens seeking advice. It might be argued that the same line of argument is available in this case, and that citizens too are benignly deceived when they turn to experts for making an informed decision. However, there is an asymmetry between the two cases. Indeed, in this second case citizens do not participate in the form of life of science. Consequently, contrary to scientists, citizens do not enjoy any good from being deceived. Indeed, the only good that they could enjoy would be an epistemic one, that of being correctly informed, since this is the goal at which they aim. But, according to Collins and Evans, this is a myth. It follows, therefore, that in the case of citizens deception cannot be good. It is deception pure and simple.

4. *A Pragmatist Theory of Expertise*

The last remarks were intended to show that Collins and Evans's moral defense of science leads to morally unacceptable conclusions. If our argument is correct, therefore, elective modernism is unsatisfactory by its own standards. It is a kind of double-truth theory which ushers in in a strong form of elitism and mass-manipulation[3]. More relevantly for our purposes, it also entails the impossibility of any relation between experts and laypeople because of the illusoriness of the ground on which they would enter

[2] It should be noted that Collins and Evans's views on this issue are more complicated than this. They acknowledge that it is possible – though quite rare – for a scientist (natural as well as social) to be aware that Truth cannot be achieved, and nonetheless to continue to play the game of science, which revolves around the search for truth. Collins and Evans call "owls" those scientists who have this capacity of double vision. So, properly speaking, self-deception is not a necessary condition for being a scientist. However, while working as scientist, an "owl" cannot adopt the reflective attitude which reveals that Truth does not exist; it is only when she reflects on her activity that she can reach that conclusion.

[3] On this point see Barrotta (2018: 169 ff.).

into contact. It seems highly plausible, indeed, that if lay people were to know that – no matter how well acquainted with the subject-matter of their research scientific experts might be – these so-called experts have no epistemic warrant for their opinions, they would stop asking them for advice. Consequently, the problem of expertise would fade away. Either expertise is an epistemic notion or it is a deceit.

The pragmatist theory of expertise that we are going to outline in the following pages starts by acknowledging precisely this fact: expertise is an essentially epistemic notion. Citizens turn to experts on the exclusive belief that the latter's opinion is warranted, and that by acting on the experts' advice they have the greatest chance of reaching the desired goal. Clearly, citizens do not possess the epistemic resources to assess the validity of the responses given by scientific experts. This is an *a priori* condition: indeed, if citizens were able to acquire enough expertise to peer-evaluate experts' opinions, they would cease to be citizens and would become experts in turn. Once again, the problem of expertise would fade away.

It follows, therefore, not only that the notion of expertise is essentially epistemic, but also that it is intrinsically interwoven with the notion of trust. There is no expertise without trust: scientific experts are those who are judged trustworthy by citizens. This is the thesis that we want to articulate through our pragmatist theory of expertise.

Such a thesis is likely to look highly problematic. For instance, it may be countered by arguing that it leads to a dangerous submission of scientific expertise to the judgment of lay people. Clearly, if science were made dependent on what citizens happen to think is true, the autonomy of science would be fatally weakened. This is an unfortunate consequence which threatens the possibility of objective knowledge, and, consequently, undermines the very idea of expertise. Indeed, citizens ask experts for advice because their opinions are taken to be true in and of themselves: if scientific opinions were in need of any kind of external support – external to the body of the scientific community – they would become a matter of preferences and polls. But then this would be nothing but Second Wave approach in disguise.

Clearly, this is not what we have in mind. Quite the opposite, our proposal aims to preserve the autonomy of science, and, at the very same time, to acknowledge the relational nature of expertise. In order to see how this is possible, it is useful to first clarify the concepts that we will employ. In particular, we want to introduce a distinction between scientists and scientific experts. Though not widely accepted, this distinction is not

wholly new[4]. As we intend it, the distinction purports to highlight a difference in the way scientists operate.

We rely here on an insight shared by Collins and Evans. As said above, elective modernism is concerned with technical decision-making. Collins and Evans rightly remark that what is at stake in cases of technical decision-making is the solution of a particular problem in which social and scientific issues are inextricably entangled together. The subject-matter of technical decision-making is, therefore, a complex *imbroglio* which cannot be reduced to its scientific components. Think, for instance, of the construction of a nuclear power station in one specific locality. Clearly, many of the problems to be dealt with in planning the construction work have to do with scientific and technical issues – from the composition of the concrete, which must be not too porous, to the design of reactor containers and the assessment of the irradiation effects. However, other legitimate problems arise, such as the opportunity to build in that particular site, the economical and social consequences of that project, the political and military risks that inevitably have to be faced, the ethical concerns about the impact of that decision on future generations, and so on. All these social aspects are just as relevant for the definition of the problem as its scientific components.

What is this example intended to show? We believe that it helps to shed light on the fundamental difference between science as it is carried out in laboratory and science as it is conducted in the public space. In the former case, the subject-matter of scientific research is abstracted and idealized; in the latter case, on the contrary, the subject-matter of technical decision-making is a group of processes and events taken in their concreteness. The example also highlights the different complexity of the subject-matter: by stressing the fact that technical decision-making cannot be boiled down to its scientific and technical components, it is implied that science cannot provide the whole truth of the matter. Any threat of technocracy is thus excised.

The difference between science in laboratory and science in the public space is what we want to grasp through our distinction between scientists and scientific experts. A few considerations are worth making here. First of all, that distinction is *functional*: the very same person can be a scientist and a scientific expert, as a consequence of being engaged in different activities – respectively, scientific research and technical decision-making. This does not mean, however, that we are committed to a relational

[4] See, for instance, Grundmann (2017).

conception of what it means to be a scientist. Much of the recent debate on the nature of expertise has turned around this issue, whether expertise is a substantial or relational notion. One of the strengths of our approach is that it allows us to take the best of both worlds. Thanks to the distinction between scientists and scientific experts we are allowed to say that being a scientist is a substantial qualification: in order to become a scientist, one has to reach a certain number of educational and academic achievements, not least of which is getting an academic job. On the contrary, to be acknowledged as an expert is a relational notion, which we conceive of as based on trust.

Our proposal is in agreement with the ordinary use of the terms: while we say that being a scientist is a profession, being an expert is a status, and the attribution of such a status is context-sensitive. Indeed, to be recognized as an expert depends partly on the specific problem at stake, and partly on the background knowledge of those who turn to experts for advice. So, for instance, if I do not know anything about wine, asking a sommelier who can give me tons of information about the different methods of production of wine is much less effective than asking to a wine shop assistant who can provide some educated guidance. At the end of the day, time and intellectual effort matter when one has to make a decision.

In more precise terms, being a scientist is a *necessary* condition for being a scientific expert, but it is not a *sufficient* condition. A scientist turns into a scientific expert when she is asked to participate in technical decision-making. However, as pragmatists never tire of pointing out, the application of a body of knowledge is not epistemically neutral: it raises new problems, and asks for different solutions. The problems that a scientific expert has to face are different from the ones that she faces when she works as a scientist. This is partly due to the fact that any concrete case presents some specific features which must be taken into account, and which cannot be derived from the body of knowledge already available (Barrotta & Montuschi 2018). In addition, there may well be reservoirs of information that are not formulated in scientific language, but nonetheless prove to be reliable and valuable[5] (Wynne 1996). Finally, as has been

[5] It is worth noting that the acknowledgment of the existence of reservoirs of information that are possessed by lay people does not enter into conflict with our assumption that being a scientist is a *necessary* condition for being a *scientific* expert. We do not want to take a position on the issue of lay expertise since there is not enough space to provide a detailed discussion. It is sufficient to remark that we can easily make room for that concept in our account by distinguishing between different varieties of expertise.

repeatedly stressed, in the case of technical decision-making the subject-matter is made more complex by the entanglement of scientific and political issues.

We are in a better position now to clarify the conceptual import of the notion of trust. We have stressed the fact that the "grammar" of expertise is grounded on trust, and that trust is an essentially epistemic concept. No expertise without trust, therefore. Trust, however, should not be conceptualized in an unidirectional way, as going from citizens to experts. If it were, our proposal would be substantially identical with the First Wave idea of "public understanding of science". On the contrary, trust is a bidirectional relationship: it is only because they succeed in being perceived as trustworthy by citizens that scientific experts are so acknowledged. Contrary to being a scientist, which is a profession, being a scientific expert is a social status that has to be earned and maintained. Trust can be withdrawn any time.

At the very same time, however, our pragmatist account of expertise provides some strong normative criteria to evaluate the *legitimacy* of citizen dissent against the advices of scientific experts. Dissent is not legitimate when it is directed against propositions a) that are accepted by the scientific community, and b) whose content is unaffected by any social consequences in which the objects which the propositions refer to may take part.

Take, for instance, the protests against vaccines. Are they legitimate according to our approach? The point at stake is to understand what these protests take as their target. If they are directed against settled scientific facts – such as the fact that vaccines do not cause autism – then they are not legitimate since they would interfere with the proper domain of science. Trusting scientific experts means to acknowledge and respect their competence in their field of expertise: public dissent has limits, which are defined by our best method of ascertaining the truth of a proposition. On the contrary, if the reasons of the protest have to do with the opportunity to publicly finance a campaign of vaccination, then the dissent is legitimate since the subject-matter of the problem is a social issue which cannot be boiled down to its scientific components. Here, trust puts some normative constraints in the opposite direction: scientific experts, in order to earn and preserve their status, are compelled to acknowledge the right of the citizens to participate – as epistemic contributors – in technical decision-making processes.

5. *Conclusion*

The goal of this article was to criticize Collins and Evans's moral defense of the role of science in democracy, and to point out that, contrary to what they believe, the notion of scientific expertise is epistemic through and through. We have shown that devoid of its epistemic dimension, the appeal to scientific expertise turns into a form of deception of the citizens. Then, we have argued that trust should be conceived of as the backbone of scientific expertise. Our pragmatist account of expertise revolves precisely around the idea that being a scientific expert is a social status that is to be earned and preserved: scientific experts are those who are perceived as trustworthy by the citizens. Finally, we have stressed that trust is a bidirectional relationship. More precisely, trust is a normative concept which puts constraints on the kinds of behavior that citizens and scientific experts are legitimate to perform. It follows that technical decision-making is a highly dynamic and conflictual sphere, in which the struggle for reciprocal recognition goes hand in hand with the effort to find the most reliable solution to the problem at stake.

References

Barrotta P. (2018), *Scientists, Democracy and Society: A Community of Inquirers*, Springer, Berlin-New York.

Barrotta P., Montuschi E. (2018), *Expertise, Relevance and Types of Knowledge*, in «Social Epistemology», 32/6, pp. 387-396.

Collins H. (2014), *Are We All Scientific Experts Now?*, Polity Press, Cambridge.

Collins H., Evans R. (2002), *The Third Wave of Scientific Studies. Studies of Expertise and Experience*, in «Social Studies of Science», 32/2, pp. 235-296.

Collins H., Evans R. (2017), *Why Democracies Need Science*, Polity Press, Cambridge.

Grundmann R. (2017), *The Problem of Expertise in Knowledge Societies*, in «Minerva», 55, pp. 25-48.

Latour B. (1983), *Give Me a Laboratory and I Will Raise the World*, in C. Knorr, M. Mulkay (eds.), *Science Observed*, SAGE Publications, London, pp. 141-170.

Wynne B. (1996), *May the Sheep Safely Graze? A Reflexive View of the Expert-Lay Knowledge Divide*, in S. Lash, B. Szerszynski, B. Wynne (eds.), *Risk, Environment & Moderinity. Towards a New Ecology*, SAGE Publications, London, pp. 44-83.

Abstract

Collins and Evans's Third Way of Social Studies of Science is an ambitious attempt to counteract the de-legitimation of scientific experts that is going on in contemporary Western societies and which, on a theoretical level, represents an unfortunate consequence of the corrosive approach championed by many proponents of Social Studies of Science. Collins and Evans argue that the importance of science in technical decision-making should be defended on purely moral grounds, without having recourse to epistemic notions. The goal of this article is to criticize Collins and Evans's moral defense of the role of science in democracy, and to point out that, contrary to what they believe, the notion of scientific expertise is epistemic through and through. Our pragmatist account of expertise revolves around the idea that being a scientific expert is a social status that is to be earned and preserved: scientific experts are those who are perceived as trustworthy by the citizens. We argue, therefore, that trust is a bidirectional relationship. Trust is a normative concept that puts constraints on the kinds of behavior that both citizens and scientific experts are legitimate to perform.

Keywords: pragmatism; expertise; third wave of social studies of science; scientists; philosophy of competence.

Pierluigi Barrotta
Università di Pisa
pierluigi.barrotta@unipi.it

Roberto Gronda
Università di Pisa
roberto.gronda@unipi.it

The Prismatic Shape of Trust

T

Unwelcome Trust

Justin Bzovy

Introduction

In general, trust appears to be a good thing. A society which fosters and is built on trust is better than one that is not. Despite this, we sometimes reject the trust that other people place in us. This may be because we view the trust as a burden, because we feel that we are unworthy of trust, or even because we do not have the right sort of relationship with the person who trusts us, despite whether or not we are ourselves worthy of trust. Unwelcome trust typically arises when the trustor expects a specific type of action from trustee, but the trustee, for whatever reason, does not want to do what the trustor wants. The existence and importance of this phenomenon has been only hinted at in the literature on trust and trustworthiness. Special attention has been paid to whether or not certain accounts of trust or trustworthiness can explain any aspect of the phenomena at all. Typically this has taken the form of criticism. Karen Jones (1996: 9-11), for example, has argued against Annette Baier's (1986) and other *entrusting* accounts of trust, by suggesting that these sorts of accounts cannot handle the full range of cases of unwelcome trust. Carolyn McLeod (2002: 32-33) has in turn also argued against Jones' (1996) account of trust, by suggesting that her account cannot handle unwelcome trust. Despite the noted importance of unwelcome trust in such discourses, no sustained account of the phenomena itself has been developed. In this essay I develop such an account.

I argue that, in fact, multiple accounts of trust are needed to explain unwelcome trust. This is because trustees may reject either what it is they are being entrusted with, though not the trust, or the very trust itself, but not whatever they are being entrusted with.

TEORIA 2019/1 DOI: 10.4454/teoria.v39i1.55

1. *Trusting to Keep Secrets*

In this section I will begin by describing what I take to be a relatively straightforward example of unwelcome trust. I will then show what it is that various accounts of trust would imply about how to explain what aspect of the trusting relationship is being objected to in this sort of scenario. It is important to start with a simple example because it is not always clear how an account of trust can handle the phenomena, if an account can even handle it at all. In this preliminary analysis of unwelcome trust I will argue that we can divide these accounts into two separate camps. I will call these the trust-rejection and entrusted-rejection model, and will develop each model throughout the paper.

Consider a simple situation involving two friends, Veronica and Nathan. Veronica trusts Nathan to keep her secrets, but Nathan does not welcome this trust. He might not want to keep *those* secrets, he might not trust himself to keep them, or he might not believe that he bears the appropriate sort of relationship to Veronica for her to entrust her secrets to him. The particular reasons that Nathan has for rejecting Veronica's trust are not important. What is important about this example, as will become clear when we look at further examples, is that Veronica and Nathan are peers. There is also no relevant power differential obtaining between them. Now let's look at how some of the different accounts of trust would explain what it is that Nathan might be objecting to in this type of case. I will not be exploring all aspects of each account, only those relevant to explaining this case of unwelcome trust.

Annette Baier (1986) provides one of the most prominent and influential accounts of trust. According to her analysis, trust is a three-place relation involving a trustor, a trustee, and some valued thing that the trustor entrusts the trustee with. Karen Jones (2004: 4) refers to these sorts of account as a three-place analyses. According to Jones, three-place trust is no different from mere reliance. We can rely on thermometers to accurately report the temperature, but we cannot trust thermometers. We are not betrayed when a thermometer fails to do its job, but we often are when someone we trust breaks that trust. Thus, Baier's account distinguishes trust from mere reliance by taking trust as a special kind of reliance: reliance on another's goodwill (1986: 234). Now, though separating trust from reliance is an important task, it will not play a role here in the argument. The cases I will be considering will uncontroversially be instances of trust rather than reliance. What will be controversial is their status as cases of

unwelcome trust. Thus, in this case, Baier's account of trust would imply that Veronica is the trustor, Nathan is the trustee, and the valued thing that she is entrusting him with is keeping her secrets. According to this analysis of trust then, Nathan is not objecting to the trust *per se*, but the thing that Veronica is entrusting him with. Nathan does not want to keep Veronica's secrets.

Karen Jones (1996) offers another prominent account of trust. According to Jones, «to trust someone is to have an attitude of optimism about her goodwill and to have the confident expectation that, when the need arises, the one trusted will be directly and favorably moved by the thought that you are counting on her» (Jones 1996: 5-6). For our purposes, the key part of her account of trust involves the aspect of a confident expectation. According to her analysis, Nathan is rejecting Veronica's confident expectation that he will directly and favourably be moved by the thought that she is counting on him. In fact, this is what she herself says about a similar sort of example, though she does note that most cases of unwelcome trust will be cases where what is entrusted is rejected (Jones 1996: 9-11).

Philip Pettit (1995) offers what he calls a trust-responsive account of trust. His account is meant to cover situations wherein trustors don't have good reasons for believing that the particular trustee is trustworthy. Even without these reasons people can decide to trust if they have reason to assume that the trustee desires to be held in regard (Pettit 1995: 219). Being trusted, and being considered trustworthy, is a good thing, and the trustor will expect her trust to motivate trustees. The key part of Pettit's account for our purposes is his focus on motivation. According to his analysis, Nathan would be objecting to Veronica's trust because he is rejecting her intention to motivate him to keep her secrets that is part and parcel with her trust. That is, Nathan is objecting to Veronica's intention to motivate him.

For our purposes these three views fall into two camps. On Baier's entrusting model it is the secrets entrusted that are unwelcome and not the trust *per se*; on Pettit's and Jones' coercion models it is the trust itself that is objected to and not the secrets being entrusted *per se*. On the trust-rejection model, there are many different ways of explaining what trust is, and how it differs from notions like reliance. In terms of how we analyze unwelcome trust, these different ways of cashing out trust would simply provide further differentiations. Let's now move to a more complex example.

2. *Trusting in Professional Relationships*

In this section I will consider how the trust-rejection and entrusting-rejection models handle an example of unwelcome trust that I will develop that may hold in professional relationships. Although I use this sort of example, I believe the case will also hold in other relationships where there might be some discrepancy between how the two agents view their respective relationships. I have chosen examples from professional relationships because they are clear, plausible, and hopefully uncontroversial. I will begin by describing some situations where trust can be unwelcome in professional relationships. I will then argue that the entrusting rejection model provides the correct description of what is being objected to in these sorts of situations despite first appearances.

The sort of case that I want to consider for this section comes from reflecting on an example of a teacher rejecting the trust of her student provided by Carolyn McLeod. I quote her example in full:

> [I]f a student trusts his teacher to be emotionally supportive in the way that a parent would be, but the teacher does not (nor does she want to) think of her relationship with the student as being like that, the student's trust would be unwanted (McLeod 2002: 33).

Similar examples arise in the context of health care. A medical practitioner may not welcome the trust of her patients, when, for example, her patients trust her to make house calls or perform unnecessary procedures (McLeod 2004: 189-190). Likewise, if a patient trusts a medical practitioner to be emotionally attentive to her needs as a lover would be, and the practitioner does not think of her relationship with the patient in this way, the patient's trust would also be unwelcome (McLeod 2004: 190). What is important to note about these sorts of examples is that they, unlike the example involving secret keeping, which was between peers, involve a distinct power differential, and also include a professional element. These differences put constraints on the forms of trust that ought to obtain between the two agents. I will now show what it is that the teacher (similarly for the medical practitioner) objects to on the three accounts of trust under consideration.

Let's first consider how the entrusting-rejection model would handle this sort of case. According to this model, the teacher is objecting to the specific task of being emotionally supportive like a parent would be toward the student. But what exactly is the student «entrusting» the teacher

with? There is no obvious good or valued thing that is being entrusted. The student is not giving anything over to the teacher to look after. Thus, this model cannot adequately deal with this sort of case. However, there is a way of avoiding this initial problem, as I will discuss below, if we are less strict about the sorts of things we allow to be entrusted. Before I develop the entrusting-rejection model further, I will explicate how the two different versions of the trust-rejection model would handle this sort of case.

Recall that on the trust-rejection model it is the trust, not the entrusted thing that is being rejected. Thus, the teacher is rejecting the trust that the student is offering, not the specific thing that is being entrusted, because there is no specific entrusted thing. According to Pettit's (1995) account, the teacher is objecting because the student intends to motivate the teacher to perform an expected action, being emotionally supportive in the way a parent would, by trusting the teacher. Similarly, according Jones' (1996), there is an element of coercion. The teacher is rejecting the student's confident expectation that the teacher will directly and favourably be moved by the thought that the student is counting on them to behave as a parent would toward the student (Jones 1996: 11). Without either the intention to motivate, or the confident expectation, these would not constitute cases of trust, and so, if this is a case of trust, this must be what the medical practitioner or teacher is rejecting.

Which of these two models better fits the case? On first inspection, the trust-rejection model seems to more adequately explain the scenario than the entrusting-rejection model. It seems more plausible to say that the student is coercing the teacher to behave as the student wants the teacher to behave, and that this is what the teacher is objecting to. It does not seem correct to say that the teacher is objecting to some specific entrusted thing, because there is nothing specific being entrusted. Despite appearances, I think it is possible to say something further. I will now do this by building on McLeod's (2002; 2004) treatment of unwelcome trust.

McLeod (2004: 189-190) argues that unwelcome trust ('unwanted trust') often occurs when there is a mismatch amongst *relationship-specific commitments*. What would this imply about our present case? It would imply that, from the teacher's point of view, she has a relationship-specific commitment to behave professionally as a teacher would toward her students, and not in the way a mother would to her daughter, or in any other way that would conflict with the professional nature of the relationship. From the student's point of view, on the other hand, the student believes that the teacher has a relationship-specific commitment to act as more than just a

mere teacher. In other words, the student believes that the relationship she has with her teacher is closer to that of the relationship she might have with a parent. Of course, in an actual case, the student might not explicitly be aware that they have this sort of belief. The student may simply be confused about what to expect from this teacher, or confused about what one ought to expect from a teacher in general. Talking about this as a belief that the student has is simply a useful way of cashing out what is going on in this example[1]. The student and the teacher have a mismatch, because each views their relationship differently. According to this view, the teacher then is objecting to the way the student perceives their relationship.

McLeod's description does not, at first glance, go against my reconstruction of the trust-rejection model of unwelcome trust. However, I will argue that this is not the case. McLeod's position is best understood as a more sophisticated version of the entrusting-rejection model. I mentioned above that Baier's account has a problem with unwelcome trust in professional relationships, because it is not clear what a student is entrusting the teacher with. One possible solution, as I mentioned earlier, was being vaguer about how we construe the notion of entrusting a specific good. I will now argue that McLeod's understanding of relationship-specific commitments has provided us with a way of solving this problem for the entrusting-model.

Consider again that in this sort of example there is a clear discrepancy between how the teacher and the student view their relationship. The teacher takes the relationship to be a professional one, whereas the student sees the relationship as something more personal. The student believes that the relationship she has with her teacher is more like the one she would have with a parent. In trusting the teacher to be emotionally supportive in the way that a parent would be, the student is, to use Baier's terminology, entrusting the teacher with this type of relationship. The teacher then is rejecting this sort of relationship that the student is entrusting her with.

One might object by arguing that the notion of entrusting is now being stretched too far. We cannot seriously make sense of someone entrusting a specific type of relationship to another person. However, I am reformulating McLeod's account in this language to draw an important parallel. By drawing this parallel we can see what the difference is between the entrusting-rejection model of unwelcome trust and the trust-rejection model.

[1] Thanks to Esther Rosario for pointing this out.

According to the latter, we would say that the teacher is objecting to the coercive aspect of the student's trust. In trusting the teacher to be emotionally supportive in the way that a parent is, the student is trying to coerce the teacher to behave this way. According to the entrusting rejection-model the teacher is objecting to the type of relationship that the student is entrusting to the teacher.

I will now argue that the entrusting-rejection model offers the correct explanation of what the medical practitioner or teacher is rejecting in the sort of professional example of unwelcome trust I am examining. Consider what would happen if the teacher (or medical practitioner) explains why they do not want to be trusted in the way that the student (or patient) is trusting them. The teacher would explain to the student that this is not the way the teacher perceives the relationship, or the way the relationship ought to be perceived. The teacher would say that student's expectations for the teacher to be emotionally attentive are, so to speak, 'out of line,' considering the normal professional relationship that teachers ought to have with their students. On the trust-rejection account the teacher would have to say that the student should not be coercing (or intending to motivate) the teacher to do things she does not want to do. It is not coercion (the student's intention to motivate the teacher, or the student's confident expectation that the teacher will directly and favourably be moved by student's trust) that the teacher is objecting to simpliciter, but the specific content of what the student is trusting the teacher to do. The teacher would (one hopes) want to act in a professional manner no matter what the student would expect of them.

One might object and offer the following defense of the trust-rejection model[2]. According to this model, we can say that the teacher is objecting to the trust, because the teacher perceives this trust as coercive, as an attempt to change the relationship. That is, the trust that the student is bestowing on the teacher is eliciting that the teacher ought to offer this student a special relationship. In response, we can say that this objection gets the order of explanation wrong. The teacher here has a duty to maintain a professional relationship with her students, and she should not accept this sort of relationship under any circumstances. It is not the coercive nature of the trust that she is objecting to, but the relationship. This response becomes even clearer if we consider a situation where the student simply does not understand the nature of what the relationship is supposed to be.

[2] Thanks to Rob Shaver for this objection.

If the teacher knows this, she will simply point out that they cannot have that sort of relationship, and the student, now informed, will hopefully stop trusting the teacher in this way. In sum, this sort of example provides us with a case where the entrusting-rejection model provides the explanation to the detriment of the trust-rejection model. Nevertheless, this does not mean that the trust rejection-account of unwelcome trust is not entirely invaluable. In the next section I will show the sort of scenario where the reverse holds.

3. *Therapeutic Trust*

In this section I show that there are examples of unwelcome trust that the entrusting-model cannot handle, but that the trust-rejection model can. In order to do this, I will first develop a third sort of example of unwelcome trust that involves what is often called «therapeutic trust.» Before I develop this example, I will give a brief explanation of therapeutic trust.

If a person trusts another therapeutically, then that person «aims at increasing the trustworthiness of the person in whom it is placed» (Horsburgh 1960: 346). Unlike most cases of trust, the optimism that is typically present is lacking in this situation. Horsburgh importantly notes that: «Therapeutic trust in the full sense requires that the person trusted should be aware of the reasons for the trust which is placed in him» (Horsburgh 1960: 346). Parents often trust their teenagers with the family car or with taking care of the house while they are away with the belief (or the hope) that they will become more responsible if treated as if they are trustworthy, even if they have proven to be untrustworthy in the past (McGeer 2008: 241). Therapeutic trust is often considered to carry a normative component that non-therapeutic trust lacks, though some argue that all forms of trust carry this component (e.g., Walker 2006: 79-80). That is, when someone trusts therapeutically, the trustor is implying that the trustee should do what she is trusted to do. When parents trust their untrustworthy children with the car after they have crashed it several times, they are giving the impression that the child ought to take care of the car, even if they are not optimistic about what their child might do as they would be if their child had proven herself trustworthy.

Let us imagine a child, Sally, rejecting the therapeutic trust of her parent, Anne. Anne trusts Sally, her oldest child, to look after her younger children, Billy and Connie. Anne wants to engender trustworthy behaviour

in Sally by offering this opportunity to prove whether or not Sally can be trusted. I will leave it open as to whether Sally has, in fact, proven herself to be untrustworthy in the past, because this is irrelevant. In this sort of situation Sally might want to reject Anne's trust for several reasons. However, for the sake of argument, I am going to assume that Sally rejects this trust because she doesn't want to grow up and be responsible for Billy and Connie, and that Sally interprets Anne's trust as implying that she ought to take on this new responsibility. But how do we cash out Sally's reasons for the rejection of Anne's therapeutic trust? I will start by explaining how each model of unwelcome trust would explain this sort of case. I will then argue that the trust-rejection model offers the correct description of the case.

The entrusting-rejection model would explain the situation in the following way. There is a mismatch between how Anne and Sally view their relationship. Anne believes that their relationship is changing. By trusting Sally with babysitting duties, she takes their relationship as different than the relationships she has with Billy and Connie. In Anne's mind Sally is (or is becoming) more grown-up, more trustworthy. Anne expects Sally to interpret their relationship in a similar fashion. Anne expects Sally to view their relationship as that between a more-responsible child and a parent, rather than the relationship she has with a less-responsible child, like either Billy or Connie (or like the relationship that Sally had with Anne in the past). Sally, on the other hand, believes that she is still a kid that does not have more grown-up responsibilities, like looking after Billy and Connie. Sally wants to have the same relationship with Anne that Billy and Connie have with her. In other words, Sally is objecting to the more mature relationship that Anne is entrusting her with.

The trust-rejection model, on the other hand, can offer an alternative explanation of the situation.

According to this model, Anne is trying to coerce Sally in to taking on more responsibilities. According to Pettit's account, Sally is objecting because Anne intends to motivate Sally to babysit Billy and Connie by trusting her. Similarly, according to Jones' account, Sally is rejecting Anne's confident expectation that her daughter will directly and favourably be moved by the thought that she is counting on Sally.

I will argue that the trust-rejection model offers the correct explanation. Consider the sorts of things that Sally would say in conversation with Anne if she refused to accept the babysitting responsibilities. Sally would at first say that she simply did not want to babysit Billy and Connie. If Anne asked her to explain this further she would say that she wanted to

be treated the same as Billy and Connie. If Anne protested, and explained the motivation behind her therapeutic trust, then Sally would still refuse, and claim that she did not want to grow up, and she did not want her mother to make her grow up. Sally is not objecting to the changed relationship she is being entrusted with, she is objecting to the fact that Anne trusts her in order to change the nature of their relationship. Anne trusts her to help her grow up.

One might object that I have concocted this example to fit more closely with the trust-rejection model of unwelcome trust. But there is nothing implausible with my concoction, especially with regard to the exchange between Anne and Sally. Thus, in at least this type of case, the trust-rejection model is required to explain why the trust is unwelcome. Thus, due to the cases I have presented, and the accounts of trust I have considered, we need more than one account to deal with the phenomena.

4. *Rejecting the Stance*

One might object to the disjunctive account of trust I have offered here by suggesting a third possible model of what a trustee might be objecting to in cases of unwelcome trust. In this section I will consider whether or not such an objection can be raised by way of Richard Holton's (1994) participant-stance account of trust. I will first describe his account, and then describe a case of unwelcome trust that his account can explain, but no other can.

According to Holton, whenever we feel resentment, gratitude, or betrayal towards people when they act in certain ways, we are taking a participant stance toward them. A participant stance is what Holton calls Strawson's (1962) participant attitude. According to Holton, trust differs from mere reliance in that it involves a participant stance toward the trustee. Trust is like reliance, but trust involves a readiness to react to the trustee's actions or in-actions in a particular way. If Veronica trusts Nathan to keep her secrets, and Nathan fails to do this, because he drank too much, she will feel betrayal. If, on the other hand, Nathan keeps her secrets and beats a lie-detector, she may feel gratitude toward him.

Consider a scenario involving a healthcare aid, Jesse, and a physically and cognitively disabled client, Darcy. Darcy is capable of feeding himself, but often, seemingly on purpose, he spills his food on the floor if no one is engaging with him in a particular way. Due to the way Darcy reacts

when confronted about throwing his food on the floor, it is clear that he doesn't want to do this. He has also indicated that he does not want to have someone feed him: he wants to be able to feed himself. It also seems to be clear, from other encounters with him, that he doesn't want to be supervised all of the time. In this scenario, one might argue, it is the stance that Jesse is taking toward Darcy that he is objecting to, not the trust, nor the thing that is entrusted.

One might respond by denying the plausibility of this scenario. As a cognitively disabled agent, Darcy simply does not have the cognitive machinery available to reject, or even acknowledge that Jesse is trusting him. Further, because of this, Jesse is just wrong to even trust Darcy in the first place. A worry with this line of response is that Jesse trusting Darcy bears a strong resemblance to Anne trusting her oldest child Sally. If we are correct in saying that Jesse is mistaken to trust Darcy at all, then, so the worry goes, we should also say that Anne is mistaken in trusting Sally. Now of course we might want to respond to this worry and point out an asymmetry between the two cases.

Another response is to just say that even if we are correct to say that Darcy is rejecting the stance that Jesse is taking towards him, this is really another way of describing her trust. This would then place Holton's participant-stance account alongside Jones' and Pettit's versions of the trust-rejection model. This seems to be the correct way of describing this particular situation, however, if there is a very general participant stance that is to be understood as some prerequisite for the particular trust participant stance, then perhaps it's possible for someone to reject that general stance, though not the particular stance. Although it's perhaps the least intuitive of the three, we can call this the participant-stance model of unwelcome trust.

It seems clear from what we have said so far that it is very difficult to provide a general, unified account of unwelcome trust. Different accounts of trust line up with, at the very least, two different models of unwelcome trust: the entrusted-rejection model and the trust-rejection model. Depending on how we analyze Holton's participant-stance account of trust, and the case involving Jesse and Darcy there may even be three different models. However, I think it is best to simply consider his account as another species of the trust-rejection model of unwelcome trust.

5. *We're Not Trustworthy*

Another objection to this disjunctive account might seem to stem from accounts of trustworthiness. An objector may try to develop a case where the trustee is rejecting the trust of the trustor, because the trustee does not deem themselves worthy of trust. In order to respond to this objection, I will first describe a general account of trustworthiness. Following this I will develop and examine an example that purports to show that one can reject trust not because of the thing entrusted, or the trust, but because of being viewed as trustworthy.

Many consider trustworthiness an important virtue (e.g., Potter 2002). Similarly to distinguishing trust from reliance, a virtue-based account helps distinguish trustworthiness from mere reliability. On Potter's (2002) account of trustworthiness she distinguishes at least ten requirements of trustworthiness. I will not consider all of them in detail, but will describe those which are relevant to concocting a particular example of unwelcome trust wherein the trustee is rejecting being viewed as trustworthy. Two of the requirements she distinguishes I view as closely related. First, the requirement «that we respond properly to broken trust» (Potter, 2002, 28). Second, the requirement «that we deal with hurt in relationships-both the hurt we inflict on others and the hurt we experience from others-in ways that sustain connection» (Potter 2002: 28). The general requirement that I take both of these to be getting at is that trustworthy people will respond appropriately if they fail. If you trust a trustworthy person to do something important, and they fail to do it, whether intentionally or not, they will be apologetic and will not be indifferent or callous toward whatever harm has befallen the trustor.

Let's return to our example of secret keeping. Veronica trusts Nathan to keep her secrets, but Nathan rejects this trust. In this case, Nathan is rejecting the trust because he doesn't believe that he is trustworthy. He may appear trustworthy to Veronica, but he has good reasons to believe that he will not be able to keep her secrets. Nathan might, for example, know that he is a blabbermouth that cannot resist the urge to spread gossip, or that he will spill the beans when he is intoxicated. He also values his friendship with Veronica, appreciates the trust she is bestowing upon him, and would love to keep her secrets, if he only believed that he could. That is, he is not rejecting her secrets, or her trust, just the fact that she views him as trusworthy. Is this a plausible scenario?

One might respond by reconsidering the entrusting-rejection model of

unwelcome trust. Although it seems that Nathan is rejecting Veronica's trust because he believes himself untrustworthy, he is really just rejecting what it is that she is entrusting him with: her secrets. The belief that he is untrustworthy with respect to her secrets is why he is rejecting them, but it is the secrets, and not the trust, that he is rejecting. Thus, we can supplement the entrusting-rejection model by filling out different reasons agents may have for rejecting things that are entrusted. This shows us something very important about this model of unwelcome trust: it is not always the entrusted things themselves that are rejected, but sometimes the role these things play in a wider context. That is, sometimes trustees are willing to accept particular entrusted things, but only under specific circumstances. The context matters, and helps explain why the trust is rejected, but ultimately trust is rejected on the basis of what is being entrusted.

6. *Unwelcome Trust and Generic Accounts of Trust*

In this penultimate section I will explore the relationship between the results of my analysis of unwelcome trust and what are called generic accounts of trust (Walker 2006: 79-80). I will begin by describing what I take to be a generic account of trust. Following that, I will explicate what my account of unwelcome trust implies about the potential of a generic account of trust.

Walker argues that there is a generic account of trust that covers the different species of trust available. The generic account is meant to be useful in tracking and sorting differences among particular cases of trust (2006: 79-80). She focuses mainly on the way that different accounts of trust distinguish between trust and reliance. Her own attempt at developing a generic account of trust involves taking and developing Holton's participant-stance account as the generic view. According to Walker's generic account, we should merge the participant-stance account with the expectation that trustees act as they are relied upon (2006: 79-80). Now independently of her reasons for the importance of such an account, and given that we have solely focused on specific accounts of trust, how does unwelcome trust bear on the issue of a generic account of trust?

If my analysis is correct, and we have clear cases where one and only one model of unwelcome trust is required, then it seems the domain of the phenomena is disjoint. This means that a family of analyses of trust is required to explain the different sorts of cases where trust is rejected. How-

ever, since the domain is disjoint, there is no way of providing a unified, or generic account of unwelcome trust. This in turn, gives an argument against the possibility of a generic account of trust. However, it should be stressed, my analysis does speak to the importance of having many different accounts of trust, even if they are ultimately non-unifiable in the way that Walker's account suggests.

7. *Conclusion*

In this essay I have provided a sustained analysis of unwelcome trust. I have done so by reconstructing some general views of trust, in order that they may be shown to offer up explanations for why we sometimes reject the trust of other people. I drew two conclusions from this analysis. First, accounts of trust generally fall into two categories: what I have called the trust-rejection model and the entrusting-rejection model of unwelcome trust. There are of course finer grained distinctions to be made between each account, especially in terms of how they understand trust, but each fall into two distinct models of *unwelcome* trust. Second, both models of unwelcome trust are required to explain the phenomena. The example of unwelcome trust in professional relationships yields a situation where the entrusting-rejection model provides the correct explanation of why the trust is rejected, and the example of unwelcome therapeutic trust is an example where the trust-rejection model provides the correct explanation. I have also considered two different objections to this disjunctive account of unwelcome trust. One might think that we need to develop other models of unwelcome trust that are based on either participant-stance accounts of trust, or on virtue theories of trustworthiness. I have argued that the participant-stance account can be accommodated under the trust-rejection model, and that virtue theories of trustworthiness add an important enrichment to the entrusting-rejection model of unwelcome trust. I have also shown how this analysis of unwelcome trust bears on the possibility of developing a unified, generic account of trust. My analysis suggests that this is not possible, but that a non-unified family of trust accounts is required.

References

Baier A. (1986), *Trust and antitrust*, in «Ethics», 96 (2), pp. 231-260.

Holton R. (1994), *Deciding to trust, coming to believe*, in «Australasian Journal of Philosophy», 72 (1), pp. 63-76.

Horsburgh H.J.N. (1960), *The ethics of trust*, in «Philosophical Quarterly», 10 (41), pp. 343-354.

Jones K. (1996), *Trust as an affective attitude*, in «Ethics», 107 (1), pp. 4-25.

Jones K. (2004), *Trust and terror*, in P. DesAutels, M.U. Walker (eds), *Moral Psychology: Feminist Ethics and Social Theory*, Rowman & Littlefield, Oxford, pp. 3-18.

McGeer V. (2008), *Trust, hope and empowerment*, in «Australasian Journal of Philosophy», 86 (2), pp. 237-254.

McLeod C. (2002), *Self-Trust and Reproductive Autonomy*, MIT Press, Cambridge

McLeod C. (2004), *Understanding trust*, in F. Baylis, J. Downie, B. Hoffmaster, S. Sherwin (eds), *Health Care Ethics in Canada*, Harcourt Brace, Toronto, pp. 186-192.

Pettit P. (1995), *The cunning of trust*, in «Philosophy and Public Affairs», 24 (3), pp. 202-225.

Potter N.N. (2002), *How Can I Be Trusted? A Virtue Theory of Trustworthiness*, Rowman & Littlefield, Oxford.

Strawson P.F. (1962), *Freedom and resentment*, in «Proceedings of the British Academy», 48, pp. 1-25.

Walker M.U. (2006), *Moral Repair: Reconstructing Moral Relations After Wrongdoing*, Cambridge, New York.

Abstract

An account of trust or trustworthiness must also explain what is known as unwelcome or unwanted trust (Jones 1996; McLeod 2002, 2004). Unwelcome trust typically arises when the trustor expects a specific type of action from trustee, but the trustee, for whatever reason, does not want to do what the trustor wants. The existence of unwelcome trust raises a difficult question for any account of trust or trustworthiness. Which accounts of trust can best explain unwelcome trust? I show first how different accounts of trust and trustworthiness imply that we need two different models of unwelcome trust. The entrusting-rejection model explains unwelcome trust as a mismatch between how the agents perceive their relationship (Baier 1986; McLeod 2002,

2004). The sort of relationship that is being entrusted to the trustee is what is being rejected. The trust-rejection model of unwelcome trust, on the other hand, sees it as a matter of perceived coercion. According to this view, the trust itself is rejected, not whatever it is that is being entrusted (Jones 1996; Pettit 1995). I will argue, by way of some key examples, that unwelcome trust fits neither view, and that a disjunctive or generic account of trust is required (Walker 2006). I close by defending this thesis against two objections.

Keywords: unwelcome trust; trustworthiness; trust-rejection model; entrusting-rejection model.

Justin Bzovy
University of Alberta
Concordia University of Edmonton
McEwan University
jbzovy@gmail.com

T

A Theory of Epistemic Trust and Testimony

George Christopoulos

1. *Explicatio Terminorum & Preamble*

At least two kinds of justification can be distinguished: argumentative justification and entitlements. Argumentative justification represents the ability for a subject to *articulate* arguments for the truth of a proposition and this argument is supported by *reasons available in the cognitive repertoire of the subject*. Let us call argumentative justification a subject possesses for uptaking testimony their proprietary justification. Entitlements, on the other hand, state *a right to rely on a given cognitive practice*. Entitlements do not need to be understood or accessible to the cognizer and so entitlements are the externalist analogue of justification (Burge 1993).

The process of believing the content of a proposition presented through testimony, and thereby forming a new belief, is the *uptake of testimony*. I use the term epistemic *warrant* generally, as a positive epistemic evaluation, which includes both internalist and externalist analogues e.g. justification and entitlement. These points will come up in the discussion of my theory of trust in § 3.

1.1. *Epistemic Subject Stakes Matter*

What features prominently in my theory, and has the most explanatory power, is the influence of the epistemic subject's stakes. For a hearer, having high stakes in a testimony means perceiving the truth or falsity of the content of a testimony to have an important bearing on one's life. What is at stake is some good that is contingent on having a true belief about the subject matter of the testimony or getting it right.

 DOI: 10.4454/teoria.v39i1.56

The way I use the notion of stakes is internalist or subjective e.g. a perceived good that one has mental access to. However, the concept of stakes can conceivably be cast in externalist terms as well. When I refer to a subject's stakes I mean their subjective stakes, what matters to the individual, their perceived or subjective evaluation of what is important to them, regardless of what might be "objectively" reasonable or moral (if there is such a thing).

I will claim that stakes influence the subjects' perceptions of the epistemic environment, which in turn affects the epistemic justification by strengthening or weakening the evidentiary standards for justified uptake. Upon inspection, other theories will appear more dubious than mine for what we need them to do.

1.2. *Two Camps in the Epistemology of Testimony*

Reductionism holds that justification of testimonial uptake depends on whether agents can give (non-testimonial) positive reasons for why they accept a testimony, other than simply that they received testimony. Non-reductionism holds that testimonial justification is epistemically *basic*, typically including some prima facie right or entitlement to accept testimony unless there are stronger reasons not to. That is, receiving testimony is itself a sufficient positive reason for accepting the claim.

1.2.1. *Reductionism*[1]

Two defining commitments of reductionism are the Positive Reasons (PR) thesis and Reduction Thesis. We can distinguish the Positive Reasons condition as being necessary and sufficient for warrant/justification (PR-N&S) or simply necessary (PR-N). The real strength of reductionism is that it is better equipped to deal with situations where getting truth is very important. Reductionist uptake principles are likelier to refrain from attributing justification to situations where uptake would not be justified because they are more cautious. But the cost of being too cautious is potentially missing out on some justified uptake elsewhere. Nonetheless, this might be a small price to pay. Therefore, reductionism is here to stay, at least the part that can deal with problematic cases.

[1] Contemporary philosophers defending various versions of reductionism include: Adler (1994), Fricker (1994; 1995; 2006), Lipton (1998), Lyons (1997), Mackie (1970), Shogenji (2006) and Van Cleve (2006).

1.2.1.1. *Over-intellectualizing objection*

Reductionism over-intellectualizes *most* instances of testimony by requiring positive reasons on behalf of the hearer for uptake to be justified. *Most* instances of testimony are mundane, and the uptake of those can be said to be justified without the presence of positive reasons on behalf of the hearer. One does not have to be very imaginative here to find a slew of claims that are routinely justifiably uptaken without additional positive reasons, let alone cognitively accessible ones e.g. when asking someone waiting at the bus stop whether the bus has passed yet or not. One does not need non-testimonial positive reasons beyond the speaker's testimony for this sort of uptake to be justified. In fact, even if the speaker were lying, the hearer is still justified in uptaking the belief (though it would not be an instance of knowledge) and could not be faulted for accepting it even though it happens to be false[2]. Therefore, reductionism is over-intellectualizing most instances of testimony, namely the low-stakes and mundane, which form the majority of the body of our testimonial beliefs.

1.2.1.2. *Skepticism objection*

Much of what we come to know (or believe) comes through testimony: what we did as toddlers, facts of history, the fact that we were born, or that any country we haven't visited in fact exists. The skepticism charge is the natural follow up of the previous objection. If we are not generally disposed to provide independent confirmation of testimony, then one will find it to be a difficult or impossible standard even in mundane cases and thereby risk falling into skepticism about most things. In this way, the mundane and everyday testimonial uptake come under threat, but that is an unacceptable and unintuitive result. If one wants a theory of testimony able to account for the general acceptance of everyday testimony (or natural testimony as "tellings more generally") reductionism will seem too restrictive and unintuitive. These concerns are in part why non-reductionists have appealed to a general entitlement for accepting testimony which is more congruent with our lived experience than inordinate skepticism about testimony.

[2] Notice that how easily we attribute justification to the subject depends on their stakes. If they had an interview for their dream job and being late disqualified them from the opportunity, they could be very much faulted for so readily accepting or relying on testimony. In such a high-stakes case, they may not be justified in uptaking the testimony after all.

1.2.2. *Non-Reductionism*[3]

Non-reductionists hold that testimonial justification is prima facie justified, without appealing to a reduction to more basic sources. Important contentions of N-R views are that justification of testimony is a basic epistemic source, and the hearer has an entitlement or presumptive right to accept it. The presence of positive reasons outside of receiving testimony is not necessary to be justified in accepting the testimony; just the absence of negative reasons (defeaters) for believing testimony.

Promising aspects are that it does not threaten to force us into a generalized skepticism about testimony, or risk downplaying the importance of testimony as a prevalent epistemic source of knowledge. It conforms with epistemic intuitions about the great majority of testimonial cases and it does not unduly over-intellectualize testimony.

1.2.2.1. *Gullibility objection*

The damning objection against N-R is that it too often leads to gullibility. It comes as no surprise that people may purposely lie or deceive us, nor that testifiers may believe they are being truthful when they are mistaken. The criticism is that non-reductionism cannot deal with those cases as well because they are committed to a prima facie entitlement to accept the testimony of others. This leads Fricker (1994) to argue that we cannot square non-gullibility with one of non-reductionism's main thesis: that there is some entitlement or presumptive right to accept testimony. Because N-R necessarily leads to gullibility, and gullibility cannot be allowed in good epistemic conscience, it follows N-R cannot be allowed in good epistemic conscience (Fricker 1994).

1.2.3. *Two desiderata: between routine acceptance & reason-based rejection*

At least two desiderata must be preserved for a good epistemological theory of testimony: that it neither leads the epistemic subject towards gullibility on one hand nor general skepticism about testimonial uptake on the other. I believe the camps of reductionism and non-reductionism each capture one essence of these two desirables. Hybrid views are uniquely poised to capture both. But, as I will now show, in an attempt to safeguard against the gullibility-styled charges, the hybrid views I consider overreact and go too far in the other direction, threatening to become skeptical or

[3] Contemporary defenders of N-R include: Audi (1997), Burge (1993), Coady (1992), Origgi (2004), Hinchman (2005), and Perrine (2014).

unintuitive and out of touch with everyday experience. My hybrid view promises to be more intuitive and to strike a better balance between reductionism and non-reductionism.

2. *Other hybrid views*

At least three views have explicitly called themselves hybrids, notably Lackey's (2008), Pritchard's (2006) and Faulkner's (2011). All subscribe to versions of the Positive Reasons Thesis (PR-N or PR-N&S) which I group under heading PR-N-Always: they have an underlying commitment that cognitively accessible, non-testimonial positive reasons *are always required* on behalf of the hearer for warranted testimonial uptake. This amounts to an explicitly reductionist principle, regardless of the other alleged non-reductionist elements. Faulkner's hybrid view has earned a special mention because he carves a role for trust in his theory of testimony as a reason giving capacity, which partly insulates him from the criticism.

2.1. *Faulkner, testimony and trust*

Faulkner critique of reductionism is that it is too restrictive in what can satisfy the "reasons" requirement because it «fails to recognize how trust in a speaker can warrant uptake (Faulkner 2011: 53).» His critique of non-reductionism is based in the problem of cooperation, which is that there is always the possibility of deception and the potential rationality of lying and deceit. But knowledge through an attitude of trust is nonetheless possible because it puts the subject in contact with the *extended body of warrant* of a claim. This is all warrant of the speaker and the prior sources in the testimonial chain which is become available to the hearer through testimony. «In not recognizing that our warrant for the uptake of testimony can come from trust, the reductive theory over-intellectualizes our relationship to testimony» (*ivi*: 76).

Faulkner's illustrates by example the shortcomings of reductionism: a husband is told by his wife that the plane is boarding in fifteen minutes. No doubt he could produce an inductive argument to the truth of what his wife says, but this would distort and over-intellectualize his reasons for uptake which is rather simply he trusts her in this matter. Thus, Faulkner's non-skeptical response to the problem of cooperation states that cooperation can be rationalized by an attitude of trust, which provides a reason

that warrants uptake and puts the hearer in contact with the extended body of warrant (in the example the chain extends to the wife and perhaps the airport billboard coordinator). I agree with Faulkner about the shortcomings of reductionism and the merits of trust as a reason-providing attitude. At this juncture, I want to unpack Faulkner's use of trust since it plays a significant role in his theory and my own. Trusting is something we do and an attitude we have and take.

> The act of trusting is putting oneself in a position of depending on something happening or someone doing something. The attitude of trusting is then characterized as an attitude towards this dependence. […] With respect to testimony, we trust speakers to tell the truth and we trust testimony to be true, and we show this attitude of trust by accepting what we are told or what is said. And when acceptance is motivated by an attitude of trust – when it is a case of trusting – it issues in belief. The act of trusting testimony is the uptake of testimony (Faulkner 2011: 23).

Faulkner distinguishes, correctly I believe, between Affective and Predictive trust. Although both kinds of trust have expectations, the expectation is something different in the affective case, since it concerns another's reasons for acting.

> I expect you to see fact that I will be waiting for you at the restaurant as a reason to try to turn up on time. This is a normative expectation: I think you should see things this way and so should act for this reason; and if you don't do as I expect, or don't act for this reason – for instance if you find something preferable to do – then this failure will be liable to provoke my resentment. This thicker notion of trust, with its concern with the trusted party's motivations, I've called affective trust. Affective because the defeat of its constitutive expectation engenders characteristic reactive attitudes – those provoked by trust being let down-which identify the expectation as normative and not merely predictive (*ivi*: 24-25).

I take no issue with this notion of affective trust and readily tailor it to epistemic trust and employ it in my theory. Call that affective epistemic trust. However, I opt for a more nuanced version of predictive trust than Faulkner provides. He uses predictive trust as simply «depending on some outcome» (Faulkner 2011: 24). This is the sense in which we can trust clocks to be on time or thermometers to measure temperature. For the purposes of a theory of testimony, I propose to narrow our focus to only a relatively small subset of predictive trust which concerns agents. I take Cognitive Epistemic Trust (CET) to be more appropriate to capture the nuances of the epistemology of testimony. The epistemic components narrow it down to cases where one agent trusts an another for the truth and all the complex

cognitive calculations this may engender. This kind of trust is granted only after relevant facts and figures have been considered, or reputations examined, and this is more aptly captures the role of trust as an ability. This conception of predictive trust as CET highlights the cognitive part of reasoning about the subjectivity of other agents and is, therefore, better suited to a theory of testimony. CET can be understood as *interpersonal predictive trust*.

I see my view an expansion of Faulkner's. Notably, by explicitly implicating non-epistemic factors and mental lives of the epistemic subjects, as well as a greater role for the testimonial environment. The uptake principle I submit holds the need for positive reasons as *contingent* on the friendliness of an epistemic environment and is a better middle ground between gullibility and skepticism, while still preserving (and being indebted to) many of Faulkner's insights.

3. *My hybrid view*

The following is a statement of my theory of justified testimonial uptake:

H is justified in uptaking testimony that p from source S if and only if,
(1) S asserts that *p*
(2) H adequately perceives the epistemic safety of the environment
(3) And either Case I or Case II obtain

Case I:
(4) The epistemic environment is friendly;
(5) H has justified affective epistemic trust in S that *p*

Case II:
(6) The epistemic environment is unfriendly (or friendly);
(7) H has justified cognitive epistemic trust in S that *p*

Where Epistemic Friendliness of the environment is calculated:

$$\frac{\text{Perceived Safety of the Epistemic Environment}}{\text{The Epistemic Subject's stakes}}$$

Premise (2) ensures that the hearer is not overly sensitive (e.g. extreme paranoia) or underly sensitive (e.g. oblivious to *any and all* defeaters) to defeaters in the environment, ensuring that they are adequately perceptive epistemic agents. Premise (3) provides the case distinction between

friendly and unfriendly environments and allows epistemic trust to enter and to play its important role as an ability[4]. Premises (5) and (7) are explicated by my account of justified epistemic trust in the next section. Roughly, justified affective epistemic trust relies on a general entitlement to accept testimony for its justification, and represents the N-R component of my hybrid view. Justified cognitive epistemic represents the reductionist wing of the hybrid view and derives its (argumentative) justification from the additional, cognitively accessible, non-testimonial positive reasons of the hearer. Thus, we obtain the following uptake principle.

PR-N-Unfriendly Principle

The need for non-testimonial positive reasons is inversely correlated with the friendliness of the environment.

My hybrid view has been stated, and now I move to further clarify and defend it. First, I must give a more detailed account of justified affective and cognitive epistemic trust. Then, I will defend the inclusion of subject stakes and I will then conclude after considering an important objection.

3.1. *Epistemic Trust*

3.1.1. *The Epistemic Trust Condition on Testimony*

My hybrid theory entails a condition on testimony. Call it the Epistemic Trust Condition on Testimony (ETC): for an instance of testimonial uptake to be *justified*, it must have been instantiated through *epistemic trust that was justified*. Justified epistemic trust is a *necessary condition* for justified/warranted testimonial uptake. The implications are that variables which affect the justification of an instance of epistemic trust also affect the justification of the testimonial uptake thereafter.

Since ETC is implied by my theory, its falsity would undermine my view and its truth at least offer some support. If the relevant alternatives to my claim can be shown to be false, it can be reasonably concluded that justified epistemic trust is a necessary condition on justified testimony and that the ETC is true.

I will assume it must be possible for testimonial uptake to be sometimes justified as to avoid an overly skeptical response to our problem (since

[4] The necessity of such a condition emerges from reflections on the relationship between the environment and an epistemic subject's abilities. Pritchard (2006) discussed this relation in in his paper presenting his hybrid view.

granting otherwise would constitute too much of a departure from common sense). One relevant alternative challenge to the ETC is the claim that there can be instances of justified testimonial uptake, instantiated through *unjustified trust*. But this will consistently fail to produce intuitive results. Consider how incongruent it sounds to assert «H is justified in uptaking the belief that *p* based on S's testimony that *p*, but H's epistemic trust in S that *p* is unjustified». If the epistemic trust leading to testimonial uptake was unjustified, it would be a defeater of the uptake's justification. If a process is unjustified, the result will be as well, regardless of whether the ensuing belief is true or not. That is why asking a crystal ball questions to form beliefs, regardless of whether those beliefs are true (and even if those beliefs would be justified as the result of another method), leads to unjustified uptake. A bad method undermines the justification of the belief, and so unjustified epistemic trust cannot lead to justified uptake.

One might object that sometimes unjustified methods can lead to justified uptake. Let us entertain a case where S is seemingly generally untrustworthy, yet H is justified in epistemically trusting S that *p*. For example, say that S has no knowledge or expertise about anything other than automotive matters, but in that, he is widely hailed as an expert by other experts in the field. If S testifies that H's brakes need changing (*p*), H is justified in epistemically trusting S that *p* and uptaking the belief that *p*. But this objection fails, because one can be generally untrustworthy on all matters not relevant to *p*, but as long as they are trustworthy with respect to *p*, H is justified in trusting S that *p* and uptaking the belief that *p*, and so the method is not unjustified with respect to *p*.

Another alternative to the ETC is that testimonial uptake can be justified *without* the presence of justified epistemic trust at all. But one will be hard pressed to think of such examples because, intuitively, justified testimonial uptake depends crucially on something gained through epistemic trust: it puts the epistemic subject in a position to be in contact with the extended body of warrant of a claim, and this does a great deal of justificatory work for that claim. Consider the important difference between the case where I surmise (truthfully) that my neighbour D is upset (*p*) *and* believe that *p* based on nothing but my own, perhaps lucky, whims. Contrast this with the case where I form the belief that *p* based on my neighbour's partner's testimony that *p* (who derives warrant for believing that *p* from first-hand experience that *p* through, either by direct perception or from receiving testimony from their partner). The second case puts me in contact with the extended body of warrant (in this situation, the warrant of D's

partner belief that *p).* The second case is connected in some way to the truth while the first is not, and this role of tethering is played by justified epistemic trust and this would hold if the chain of warrant was longer as well.

Furthermore, consider how odd it sounds to say «H justifiably uptakes new belief *p* on the basis of S's testimony that *p*, but H does not have justified epistemic trust in S that *p*». How could it be that there is no justified epistemic trust? After all, as we have seen, epistemic trust puts the subject in contact with the extended body of warrant. If the belief is based in testimony, and uptake is to be justified, the epistemic trust in which the belief is based in must be justified. Although perhaps not logically contradictory, the utterance seems intuitively incoherent. These considerations suggest we can safely conclude justified epistemic trust is a necessary condition on justified testimonial uptake and therefore ETC is true, lending some initial support to my view.

3.1.2. *Epistemic Trust: Between Skepticism and Gullibility*

I suggested that non-negotiable desiderata for a theory of testimony are that it does not entail either general skepticism or gullibility. An integral motivation of my theory of trust is that precision tools like PR-N-Unfriendly will get the correct answers more often than brute force maneuvers of other theories. It seems commonsensical, from the onset, that a theory of testimonial uptake should not entail skepticism about most of beliefs uptaken as a result of wielding the theory. There are, of course, situations where skeptical reservations from the hearer are justified. However, when considering the group of testimonial situations as tellings more generally, it becomes obvious that skeptical reservations would be disproportionate in a majority of cases. We are told a great many things that ground much of what we take ourselves to know, and so general skepticism is an unintuitive result for a theory. My uptake principle holds a weaker version of the PR-N-Always condition; therefore, one cannot object that mine entails skepticism without also implicating the reductionism or other hybrid views even more harshly.

As for the other desiderata, even if it is granted that non-reductionism gets the correct result *in most cases* of natural testimony or tellings generally, it is not without problems. Granting that testimonial situations which require independent justification (on top of receiving testimony) may be less in number, they are often higher-than-average-stakes cases. Call this minority of cases requiring additional non-testimonial support the *problematic cases*.

This minority, however, is a *majority of the important cases* where getting it wrong can have serious consequences. That is where the gullibility objection gets its main thrust. After all, one cannot be said to be gullible for accepting mundane testimony about the weather or the speaker's favorite color. The charge of gullibility only relevantly applies to the important or problematic cases. Generally speaking, for problematic cases, stronger uptake principles (more demanding) are likelier to obtain the correct result because they appeal to something above and beyond a no-defeater condition alone. Therefore, it seems just as intuitive that a good epistemic theory of testimony should not entail gullibility about important matters either.

I readily grant that no amount of intuitiveness or expeditiousness of a theory is worth the cost getting the wrong answer in important cases. To cope with these worries about gullibility, PR-N-Always emerges as a catch-all candidate for justified epistemic uptake. But, this swings the pendulum too far the other way, because requiring blanket positive reasons for justified uptake stifle one's ability to justifiably uptake mundane knowledge through testimony, putting in danger a majority of our testimonial beliefs about people, the time, the weather and many more of life's wonderful trivialities. My view can cope with the gullibility charge no worse than reductionism or other hybrid views, since the charge of gullibility *only relevantly applies to important cases*, and those high-stakes, important cases require a reductionist uptake principle on my theory anyway. Thus, I conclude my view is no worse off than reductionism or non-reductionism in facing respective objections, and it is likely better off. It lessens the force of objectives, as we have seen, and now I will show how it preserves the best aspects of each view.

3.1.3. *Epistemic Trust and the Two Pathways*

Some authors have appealed to a dual-pathway model to preserve both aforementioned theoretical desiderata. Thagard (2005) notes that a general theory of testimony must to be able to explain how «testimony is usually accepted automatically but also how it sometimes provokes extensive reflection about the claim being made and the claimant who is making it» (Thagard 2005: 297).

My theory takes epistemic trust to be uniquely poised to fulfill that role. It posits epistemic trust has a dual-nature: two pathways, a default and a reflective pathway, where practical interests act as a trigger that shifts from one to the other. The dual nature of ET is reflected in two types of

processes, Affective (AET) and Cognitive (CET). These roughly reflect Kahneman's (2011) System 1 and System 2 pathways. The latter is more deliberate, slower and thoughtful, while the former is more intuitive, faster and emotional. Epistemic trust is CET when it involves decision-making, over a length of time, by a process which includes *consciously* weighing reasons like in rational deliberation, reflection or thinking. Epistemic trust is AET when best understood as an attitude, when it is instantiated near instantaneously, through the minimal weighing of reasons or deliberation one is conscious of (there can be reasons, but it is not required that the agent consciously employs those reasons).

3.1.4. *Justified Cognitive and Affective Epistemic Trust*

Cognitive epistemic trust is justified in a straightforwardly reductive sense. For CET to be justified, it requires additional non-testimonial positive reasons, which are accessible to the hearer. As with reductionism, the details can be filled in different ways, perhaps by appeal to reliabilism. My theory leaves the question of how to best describe reductive justification relatively open. I will not go down that relatively well-beaten path because, as noted, my theory is no worse off than one's favorite rendition of reductionism for defending against the gullibility charge, because the same details as one's favorite reductionist theory can be filled in on my view.

Justified affective trust, on the other hand, is a basic-belief forming method. It is used by infants and toddlers use to build up foundational knowledge about the world before they have the deliberative reasoning skills required by cognitive epistemic trust. Enoch & Schechter (2008) provide a solid account of how basic-belief forming methods are justified.

On their account, what explains a subject's justification in employing basic-belief forming methods such as "Inference to the Best Explanation" (IBE), Modus Ponens or relying on perception and memory is the «indispensability to a rationally required project» (Enoch & Schechter 2008: 556).

Their account is stated as such:

> A thinker is prima facie justified in employing a belief-forming method as basic if there is a project that is rationally required for the thinker such that:
>
> (i) it is possible for the thinker to successfully engage in the project by employing the method;
>
> (ii) it is impossible for the thinker to successfully engage in the project if the method is ineffective. Moreover, where clauses (i) and (ii) apply, it is in virtue of these facts that the thinker is so justified (Enoch & Schechter 2008: 556-557).

For them, a "rationally required project" is one that a rational epistemic agent must engage in. Examining the environment around us, obtaining knowledge about it and constructing a framework from which to understand it are candidates for such a project.

Indeed, building up knowledge of the world as children is arguably an exemplary candidate for a rationally required project, and affective epistemic trust is the method that, when employed, allows us to successfully engage in that project when employed. Furthermore, as infant epistemic subjects, without the relevant cognitive abilities to engage in CET or other reductionist methods, it is impossible for them to successfully engage in the project of building up knowledge of the world without that method. Even in the adult world we are often rationally required to engage in the method of AET. For example, when being trained for a new job outside of our expertise by a superior, or when reading nutritional information on a cereal box[5]. I piggyback on their account of the justification of basic-belief forming methods rather than offer additional argumentation for what I already consider an extremely plausible and intuitive account.

3.2. *Confusing Pragmatic and Epistemic Justification?*

It can be objected that my uptake principle (and perhaps my theory more generally) conflates or confuses epistemic justification with prudential or pragmatic justification. The term "pragmatic encroachment" has been used to refer to the notion that pragmatic considerations encroach on epistemic ones and one way to understand this is that there are practical conditions on knowing.

Of course, being offered a large sum of money to believe something might make you *pragmatically justified* to believe it, but obviously will not make it likelier that the claim is true. I readily grant this. But this does not refute my theory, nor pragmatic encroachment more generally. All that shows is that subjects' stakes have no bearing *on the truth of a claim*. However, it does not follow that subject stakes have no bearing on *the justification or knowledge of a claim*. Truth is but one condition on knowledge, albeit the most obvious one, and whose absence would be the most noticeable. But there is plainly more to knowledge than truth. I submit,

[5] Barring extraordinary circumstances, say we are allergic to nuts and investigating whether it contains nuts, but this case merely confirms that CET is required in that case due to higher stakes in the claims, supporting my claim.

therefore, there is no principled reason as to why knowledge cannot or should not have a fourth (or N-th) practical condition e.g. sensitivity to subject's stakes. Perhaps justification is sensitive to practical factors, which play a part in determining when it is invokable and to what degree.

At least two broad strategies have been used to support the pragmatic encroachment hypothesis. The first is an appeal to intuitions and subsequent empirical data about epistemic attributions regarding philosophical cases. This includes experimental philosophy and the results from empirical studies probing layperson and philosophers' intuitions alike (Croce & Poenicke 2017; Sripada & Stanley 2012). The second strategy is to make a theoretical case for a pragmatic condition on knowledge. That is, whatever theory of knowledge one holds (of the form JTB + x) should be supplemented by an additional condition *p* which requires sensitivity to subject interests. Many authors have argued for a practical condition on knowledge, notably Fantl & Mcgrath (1996; 2002); Hawthorne (1994); Stanley (2005).

The PR-N-Unfriendly states that stakes affect testimonial environment in a relevant way such as to influence warrant and knowledge. My view is, of course, compatible with the notion of pragmatic encroachment regarding justified testimony and justified trust.

3.2.1. *Retroactive Sensitivity: Justification and Stakes*

To support my view, and continue answering the previous objection, I want to offer a case that establishes the plausibility of the claim that there is an intuitive link between attributions of epistemic warrant and the stakes of epistemic subjects. I submit that the link is *so* strong that the attributions of warrant to a belief can change *retroactively* if stakes shift too drastically, which is something we would expect to find if my theory were true. Consider this case meant to support the idea that stakes are connected to and influence epistemic warrant and even knowledge.

Hearer H is told by reliable speaker S that carrots are safe for dogs (*p*). H is not a dog owner and is not acquainted with any dogs, and so has relatively low stakes pertaining to this claim. H trusts S that *p* and goes on a good while with this inconsequential belief. I take it that H was justified (entitled, warranted) in uptaking based on S's testimony on this matter in a way that can lead to knowledge. However, eventually, a new friend of H, call them F, entrusts H to care for a cherished dog. Friend F is in a hurry for a family emergency and leaves town without any specific canine dietary information for H.

H's stakes regarding the initial belief have now presumably gone up

and with them the justificatory evidential demands on H's trust in S that *p*. These heightened epistemic demands would require *additional reasons for sustaining the belief*. It can be said that H *no longer knows* (believes, trusts with justification), in any relevant sense, that carrots are safe for dogs. Additional positive reasons would be required to support this claim e.g. a quick search engine consultation revealing carrots are generally safe for dogs. When stakes shift drastically, there is a retroactive change in the epistemic status of the belief – H is no longer justified (entitled, warranted) to cling to the initial testimonially-based belief under the new heightened evidential pressure from having higher stakes. In the low-stakes situation, the testimonial uptake is justified because of the general friendliness of the testimonial environment (because the perceived safety of the initial testimony remains fixed, but stakes change, skewing the friendliness score).

When the requirement to justifiably uptake belief is very low like in the first case, almost any minutiae of evidence is sufficient to know, provided the belief is true. Assuming the belief is true, in the first case, simply receiving testimony is good enough evidence to qualify the belief as knowledge. But when the stakes suddenly shift, the evidential requirement to know becomes much higher because the evidential threshold on justification becomes much higher.

Imagine if the first time H were to receive the same testimony from his original friend, speaker S, after already being entrusted with the dog. In that case, justified uptake of that claim might be harder to come by. And it would seem to require something like *additional non-testimonial positive reasons*. But the fact that even the justification of *previously-held testimonial beliefs* come under threat when stakes change suggests a deep connection between epistemic justification and stakes; one preserved even after the initial moment of uptake. As we have seen from the case of the dog and the carrots, a change of stakes can change the evidential threshold for a piece of data to count as good evidence (perhaps to count as evidence at all). Though a full treatment of the practical conditions on justification and knowledge is beyond the scope of this paper, however, I hope to have sufficiently called into question the objection's appeal to the assumption that positive epistemic standings are free from non-epistemic factors. Even when those factors do not relate to the truth of the matter at all, they can still influence epistemic standings.

3.3. *Conclusion*

I have suggested there are two desiderata of a theory of testimony and argued that mine can more easily preserve both. I do not reject the view that one *can have* justified testimonial uptake or testimonial knowledge *absent any non-testimonial positive reasons*. My view even draws a principled distinction when it is possible and when it is not. At the very least then, it can be said that my theory does better against the over-intellectualization and skepticism objections.

Accounting for the full extent of epistemic subjects' mental lives in the shaping and interpretation of the epistemic environment helps perform better against the gullibility objection than non-reductionism and the other hybrids because the charge of gullibility only really applies to high-stakes cases, and my theory calls for reductionism in those cases, protecting itself in a calculated manner.

I conclude with the notion that the dual-process theory of epistemic trust I have presented can help resolve some kerfuffle in the literature on trust more generally. There are affective accounts (e.g. goodwill attitude, Baier 1986; Jones 1996), which contend that trust is primarily affective, and any sort of deliberation or cognitive component falls into the trap of being "contractual" and does not capture the "leap of faith" required to trust another. These accounts would be opposed to cognitive accounts (e.g. the expectation account of trust in Hollis, 1998) and consider them not to be instances of trust at all. Closely related is the debate whether trust is volitional. Contra Baier (1986) and Jones (1996), my theory suggests we can. at least sometimes, willfully trust another (at least in CET). These seemingly opposed views can co-exist peacefully on my dual-nature view of trust because they simply correspond to different dimensions of trust, AET and CET respectively. My theory can house both families of views and does not force us to choose between either, rather allowing us to keep the explanatory and intuitive power of both.

References

Adler J.E. (1994), *Testimony, Trust, Knowing*, in «The Journal of Philosophy», 91 (5), pp. 264-275.

Audi R. (1997), *The Place of Testimony in the Fabric of Knowledge and Justification*, in «American Philosophical Quarterly», 34, pp. 405-422.

Baier A. (1986), *Trust and Antitrust*, in «Ethics», 96 (2), pp. 231-260.

Burge T. (1993), *Content preservation*, in «The Philosophical Review», 102, pp. 457-488.

Coady C.A.J. (1992), *Testimony*, Oxford University Press, Oxford.

Croce M., Poenicke P. (2017), *Testing What's at Stake: Defending Stakes Effects for Testimony*, in «Teorema: Revista Internacional de Filosofía», 36 (3), pp. 163-183.

Enoch D., Schechter J. (2008), *How Are Basic Belief Forming Methods Justified?*, in «Philosophy and Phenomenological Research», 76 (3), pp. 566-567.

Fantl J., McGrath M. (2002), *Evidence, Pragmatics, and Justification*, in «The Philosophical Review», 111 (1), pp. 67-94.

Fantl J., McGrath M. (1996), *Knowledge in an Uncertain World*, Oxford University Press, Oxford.

Faulkner P. (2011), *Knowledge on Trust*, Oxford University Press, Oxford, pp. 24-76.

Fricker E. (1994), *Against Gullibility*, in B. Matilal, A. Chakrabarti (eds), *Knowing from Words*, Kluwer Academic Publishers, Dordrecht, pp. 125-161.

Fricker E. (1995), *Telling and Trusting: Reductionism and Anti-Reductionism in the Epistemology of Testimony*, in «Mind», 104, pp. 393-411.

Fricker E. (2006), *Varieties of Anti-Reductionism about Testimony: A Reply to Goldberg and Henderson*, in «Philosophy and Phenomenological Research», 72 (3), pp. 618-628.

Hawthorne J. (1994), *Knowledge and Lotteries*, Oxford University Press, Oxford.

Hinchman E.S. (2005), *Telling as Inviting to Trust*, in «Philosophy and Phenomenological Research», 70 (3), pp. 562-587.

Hollis M. (1998), *Trust within reason*, Cambridge University Press, Cambridge.

Jones K. (1996), *Trust as an Affective Attitude*, in «Ethics», 107 (1), pp. 4-25.

Kahneman D. (2011), *Thinking, Fast and Slow*, Farrar, Straus and Giroux, New York.

Lackey J. (2008), *Learning From Words*, Oxford University Press, Oxford.

Lipton P. (1998), *The Epistemology of Testimony*, in «Studies in History and Philosophy of Science», 29, pp. 1-31.

Lyons J. (1997), *Testimony, Induction and Folk Psychology*, in «Australasian Journal of Philosophy», 75, pp. 163-178.

Mackie J.L. (1970), *The Possibility of Innate Knowledge*, in «Proceedings of the Aristotelian Society», 70, pp. 181-196.

Origgi G. (2005), *Is Trust an Epistemological Notion?*, in «Episteme», 1 (1), pp. 61-72.

Perrine T. (2014), *In Defense of Non-Reductionism in the Epistemology of Testimony*, in «Synthese», 191 (14).

Pritchard D. (2006), *A Defence of Quasi-reductionism in the Epistemology of Testimony*, in «Philosophica», 78.

Shogenji T. (2006), *A Defense of Reductionism About Testimonial Justification of Beliefs*, in «Nous», 40, pp. 331-346.

Sripada C.S., Stanley J. (2012), *Empirical Tests of Interest-Relative Invariantism*, in «Episteme», 9 (1), pp. 3-26.

Stanley J. (2005), *Knowledge and Practical Interests*, Clarendon Press, Oxford.

Thagard P. (2005), *Testimony, Credibility, and Explanatory Coherence*, in «Erkenntnis», 63 (3), pp. 295-316.

Van Cleve J. (2006), *Reid on the Credit of Human Testimony*, in J. Lackey, E. Sosa (eds), *The Epistemology of Testimony*, Oxford University Press, Oxford, pp. 50-74.

Abstract

This essay connects the justification of trust and the justification of testimony. I provide a theory which entails that justified epistemic trust *is a necessary condition on* justified testimonial uptake. *Two important desiderata of a theory of the epistemology of testimony are that it does not lead to generalized skepticism, nor is it susceptible to gullibility about important cases. The proposed theory of testimony doubles as a theory of epistemic trust that is better than alternatives. My theory posits two kinds of Epistemic Trust (ET): Affective and Cognitive Epistemic Trust (AET and CET). I argue both processes can be justified (JAET and JCET) and both can lead to justified uptake of testimonially-based beliefs. My theory of epistemic trust distinctly carves a role for subject stakes: when they are high, the evidential justification conditions on epistemic trust become more exacting on the testimonially-based beliefs they support.*

Keywords: trust; testimony; justified epistemic trust.

George Christopoulos
Concordia University
christopoulosg6@gmail.com

T

«I Don't Trust You, You Faker!» On Trust, Reliance, and Artificial Agency

Fabio Fossa

Introduction

In Asimov's 1975 story *A Boy's Best Friend* Jimmy, a young dweller of Lunar City, instructs its robotic dog Robutt not to get out of his sight and exclaims: «I don't trust you, you faker!»[1]. Trust and distrust in robots and computers is indeed a recurring theme in many of Asimov's stories. In *Robbie*, for instance, Grace struggles to accept leaving her little daughter to the care of a robotic nanny, of which however her pupil Gloria becomes quickly fond. While discussing the matter with her husband, Grace exclaims: «I won't have my daughter entrusted to a machine – and I don't care how clever it is. It has no soul, and no one knows what it may be thinking»[2]. In *Reason*[3], Powell and Donovan wonders whether Dave, a robot that has concocted an absurd interpretation of its own condition, is to be held trustworthy. In *Point of View*[4], a smart child asks whether the Multivac, a supercomputer his father is working on, can be trusted even though sometimes it makes trivial mistakes.

As Asimov did not miss to notice, the impact of information technologies on trust relationships is deep and multifarious. The more human beings rely on technological products to accomplish their aims, the more the mediation provided by such technologies affects trust relationships and modifies their characters. Information technologies impinge on trust in at

1 I. Asimov, *The Complete Robot*, Doubleday, New York 1982, p. 4.
2 *Ivi*, p. 138.
3 *Ivi*, pp. 227-244.
4 *Ivi*, pp. 37-40.

 DOI: 10.4454/teoria.v39i1.57

least two different situations. The first situation, that of *e-trust* or *online trust*, occurs when trustors and trustees are Human Agents (HAs) who get in touch through digital platforms, mostly the internet. In general, scholars working in this field of inquiry try to shed light on «the *definition* and the *management*» of e-trust[5]. Specific problems are, for instance, how trust can be secured in digital environment[6] and what connection exists, in online contexts, between trust and reputation[7] or knowledge[8].

The second situation, which may be labelled *robotrust*[9], occurs when human trustors put trust in artificial trustees. In this case, trust relationships are not supposed to concern human actors exclusively, but to occur between human users and technological products as well. In particular, similar relationships are thought to emerge when human beings delegate tasks to autonomous technologies, such as AI systems and robots. The basic idea underlying these inquiries is that, since relationships between human beings and Artificial Agents (AAs) happen to arouse expectations of trust, it is necessary to "update" our conception thereof in order to include AAs as possible trustees. Finally, some scholars claim that trust frameworks should also be applied to the study of mutual relationships between AAs in Multi-Agent Systems (MAS)[10]. We may name this last domain *artificial trust*.

The focus of this paper is on *robotrust*, i.e., on trust relationships between human and artificial agents (HA→AA). My aim is to clarify the extent to which such relationships can be framed in terms of trust. Usually, relationships between human beings and artefacts are not supposed to imply trust, but reliance. The situation, nonetheless, appears to be opposite

[5] M. Taddeo, *Modelling Trust in Artificial Agents, a First Step Toward the Analysis of E-Trust*, in «Minds and Machines», 20 (2010), pp. 243-257, p. 244.

[6] H. Nissenbaum, *Securing Trust Online: Wisdom or Oxymoron?*, in «Boston University Law Review», 81 (2001), n. 3, pp. 635-664.

[7] T. Simpson, *E-trust and reputation*, in «Ethics and Information Technology», 13 (2011), pp. 29-38.

[8] J. Simon, *The entanglement of trust and knowledge on the web*, in «Ethics and Information Technology», 12 (2011), pp. 343-355.

[9] U. Pagallo, *Robotrust and Legal Responsibility*, in «Knowledge, Technology and Policy», 23 (2010), pp. 367-379.

[10] M. Taddeo, *Modelling Trust*, cit.; J. Buechner, H.T. Tavani, *Trust and multi-agent systems: applying the "diffuse, default model" of trust to experiments involving artificial agents*, in «Ethics and Information Technology», 13 (2011), n. 1, pp. 39-51; F.S. Grodzinsky, K.W. Miller, M.J. Wolf, *Developing artificial agents worthy of trust: "Would you buy a used car from this artificial agent?"*, in «Ethics and Information Technology», 13 (2011), pp. 17-27.

when AAs come under scrutiny. As shown in the following section, HA→AA relationships are often assumed to imply trust. I disagree with this assumption and suggest that it would be more accurate to frame *direct* HA→AA relationships in terms of reliance instead. In a word, I argue that the relationship between us and our artefacts should be interpreted in terms of reliance even when AAs are involved.

To some extent, confusion on this matter may arise since trust characterises the social milieu in which relationships between human beings and autonomous technologies occur. AAs, in fact, can also be conceived of as mediums of human actions, even though in a different way compared to how technological platforms mediate human activities in e-trust scenarios. As entities to which tasks are delegated, AAs *indirectly* mediate trust between users and other social actors involved in their design, manufacture, commercialisation, and deployment. In this sense, AAs mediate *social* trust. However, the fact that AAs mediate trust relationships between social actors does not imply that direct HA→AA relationships can or should be understood by reference to trust. The two relationships are different from each other and must not be confused.

The rest of the paper is structured as follows. Section 2 shows in what sense trust relationships between HAs and AAs have been perceived as requiring to be acknowledged rather than proved. This presupposition, however, is problematic. As argued in Section 3, in fact, trust is seldom distinguished from reliance when HA→AA relationships are discussed. Yet, since relationships between human beings and technological products are normally framed in terms of reliance, interpreting HA→AA relationships as trust relationships implies asserting that the concept of reliance does not suffice here. This, in turn, is a questionable assumption. Section 4, then, focuses on task delegation to show that AAs can only arouse expectations of reliability, i.e., not the sort of expectations that require trust to be adequately met. Therefore, placing trust in AAs appears as a form of anthropomorphism, which may result in deception and social harms. Finally, Section 5 tries to determine the extent to which AAs mediate social trust between human actors such as designers, engineers, programmers, companies, and end-users. Addressing issues related to this kind of technologically mediated trust will probably be one of the most compelling future challenges for policy makers and social institutions.

1. *A matter of fact?*

At first glance, it may seem obvious that we necessarily entertain trust relationships with AAs as soon as they enter the social stage[11]. Indeed, the social pervasiveness of trust has been strongly underscored by Luhmann[12] and repeatedly stressed ever since. Unlike traditional tools, AAs are capable of executing complex functions without supervision. This ability invests them with an ambiguous social status which falls somewhere in between that proper to things and that proper to people. Since AAs take an active part in the social organisation of work, as humans do, it is easy to see the reason why trust may seem to be required. Trust is widely recognised as one of the most fundamental elements in the organisation of complex activities. Enabling task allocation and coordination, trust allows sparing time and resources to be reassigned to new undertakings. Placing trust in others to carry out tasks aligned to a final purpose is likely to be the most effective way to face complexity and cope with multiple challenges successfully. Society cannot do without delegation, and delegation seems to require trust.

Even if the relationships between human beings and artefacts has been usually framed in terms of reliance, many authors bring trust into play when it comes to AAs. As information technologies become more fine-tuned and versatile, AAs naturally appear as suitable substitutes for human trustees. Technological products that can carry out tasks without requiring constant human oversight or intervention are great candidates for delegation. After all, robots have always been envisioned as possible substitute for human delegatees[13]. As of now, delegation of tasks to AAs is already a well-established practice in contexts as different as producing goods in factories, running driverless train systems, or providing basic costumer support. Most likely, this trend will not reverse itself soon; and the more we cooperate with AAs or let AAs operate in our place, the more it may seem sensible to think of them as trustees and of ourselves as trustors.

Besides, the rise of automation has also caused the pairing of artefacts and reliance to be surprisingly challenged. From this controversial perspective, trust is considered to be a constitutive element in relations

[11] B. Kuipers, *How can we trust a robot?*, in «Communication of the ACM», 61 (2018), n. 3, pp. 86-95.

[12] N. Luhmann, *Trust and Power*, John Wiley and Sons, Chichester 1979.

[13] N. Wiener, *The Human Use of Human Beings*, Houghton Mifflin Company, Boston 1950.

between human users and any artefact to which tasks are delegated[14]. For instance, Taddeo writes that we may trust elevators to lift us safely to our floor[15] just as, I might add, we may trust thermostats to keep the temperature constant and washing machines to wash our clothes. According to this viewpoint, delegating tasks to artefacts implies placing trust in them regardless their degree of complexity. In the context of delegation, then, trust should be primarily understood as a delegators' attitude towards any entity, be it a subject or an object, capable of carrying out specified tasks. In this sense, trust is «a *property of relations*», and thus also a property of delegation, that indicates the minimisation of «effort and commitment for the achievement» of the trustor's goal[16]. As such, trust may involve any kind of delegatee, technological products included: «As digital technologies evolve and become more refined and effective, our expectation has become an expectation to *trust* (by delegating and not supervising) them with important tasks»[17].

Other authors conceive trust as a relational dimension that may include technological products. In Coeckelbergh's opinion, for example, trust is placed on AAs mostly in light of the peculiar position they occupy in human society. «If a human-robot relation grows as a social relation», Coeckelbergh writes, «then trust is already there as a 'default' in the social relation»[18]. From this perspective, trust appears as «an emergent and/or embedded property»[19] that belongs more to delegation as a social relationship rather than to the minds of the subjects who set purposes, delegate tasks, and choose to trust. In the end, the connection between delegation and trust must be traced back to the correlation of social relationality in general – of which delegation is a case – and trust. Therefore, «in so far as robots are already part of the social and part of us, we trust them as we are

[14] B. Latour, *Where are the missing masses? The sociology of a few mundane artifacts*, in W.E. Bijker, J. Law (eds), *Shaping Technology-Building Society. Studies in Sociotechnical Change*, MIT Press, Cambridge 1982, pp. 151-180.

[15] M. Taddeo, *Defining Trust and E-trust: From Old Theories to New Problems*, in A. Mesquita (ed.), *Sociological and Philosophical Aspects of Human Interaction with Technology*, Information Science References, Hershey 2011, p. 24.

[16] M. Taddeo, *Trusting Digital Technologies Correctly*, in «Minds and Machines», 27 (2017), n. 4, pp. 565-568, p. 565.

[17] *Ivi*, p. 566. For similar considerations see also M. Taddeo, L. Floridi, *How AI can be a force for good*, in «Science», 361 (2018), n. 6404, pp. 751-752.

[18] M. Coeckelbergh, *Can We Trust Robots?*, in «Ethics and Information Technology», 14 (2012), pp. 53-60, p. 58.

[19] *Ivi*, p. 56.

already related to them»[20]. Trusting AAs does not seem to be entirely a matter of choice; rather, it appears to represent a matter of fact that necessarily follows from delegation.

Herman Tavani tackles the issue of trusting AAs in a similar vein. Building on Walker's notion of zone of default trust[21], Tavani proposes to focus on the practical contexts in which agents «come to know 'what to expect' from others and 'whom to trust'»[22]. Inside a zone of default trust, normative expectations concerning the way in which delegatees should behave emerge by disposition and, rather than concentrating on specific individuals, may diffuse on more or less undefined entities. Such expectations arise in the delegating subjects in virtue of the social situation itself, which is intrinsically determined by trust. Therefore, as long as AAs can successfully occupy the place usually reserved for human trustees, they are already «capable of being in trust relationships with human beings»[23]. From this perspective, in fact, «we can now speak of cases of various kinds that are intrinsically different, but whose common feature is that they involve a zone of default trust», so that «the concept of a zone of trust can do much of the work in assimilating a wide range of disparate cases»[24]. Trust, consequently, appears to be both a subjective disposition of a normative nature, which enables and supports task delegation, and a correlative dimensional property, which defines the features of a relational 'zone'. In sum, «HAs can enter into trust relationships with several different kinds of AAs, simply in virtue of the nature of the default and the diffuse-default zones (of trust) involved»[25]. Again, trust in AAs seems to be a simple matter of fact.

These approaches to *robotrust*, however thought-provoking they might be, are still too coarse-grained and risk masking important differences between delegation to AAs and delegation to HAs. It is certainly correct to state that AAs to which tasks are delegated occupy the social position normally reserved to human trustees. Nevertheless, once the substitution of HAs with AAs occurs, the general relationship between delegator(s) and

[20] *Ivi*, p. 59.

[21] M.U. Walker, *Moral repair: reconstructing moral relations after wrongdoing*, Cambridge University Press, Cambridge 2006.

[22] H.T. Tavani, *Levels of Trust in the Context of Machine Ethics*, in «Philosophy of Technology», 28 (2015), n. 1, pp. 75-90, p. 79.

[23] *Ivi*, p. 76.

[24] J. Buechner, H.T. Tavani, *op. cit.*, p. 42.

[25] H.T. Tavani, *op. cit.*, p. 81.

delegatee(s) requires to be equally reassessed. Even if the context in which delegation occurs is analogous, the two premises that we trust human beings, when we delegate tasks to them, and that AAs can substitute HAs as task executers, do not immediately imply that it is legitimate to place trust in artificial delegatees. The fact that, inside a zone of trust, «one simply engages in that behaviour, with little or no conscious reflection»[26] is not a justification of the behaviour itself – especially when one constitutive element of the situation in which the behaviour takes place is substituted by another that imitates it. In this case, on the contrary, it is crucial to maintain a critical focus on habitual trust attitudes to avoid deception and misplaced expectations. Even if it may feel natural to trust AAs, the question whether it makes sense to do so remains both relevant and unanswered. For this reason, it is still necessary to address the question: Can we trust AAs?

2. *Trust and Reliance*

In the context of HA→AA task delegation, the substitution of human trustees with technological products impacts significantly on the overall character of the relation. Therefore, the reasons to frame HA→AA relationships in terms of trust are not self-evident and require discussion. It thus becomes necessary to clarify whether it is accurate to transfer trust from forms of delegation that involve exclusively human beings to forms of delegation that encompass technological products as well. To this end, it must be determined first why trust is needed in HA→HA task delegation and, secondly, whether or not the same need arises in HA→AA task delegation. The point of the analysis, then, would consist in verifying whether the reasons why trust emerges in HA→HA forms of task delegation also occur in the case of HA→AA forms of task delegation. If the answer is positive, it is accurate to transfer trust from human to human-artificial contexts. Else, if the answer is negative, the concept of *robotrust* may need to be revised.

A similar enquiry is required since the theoretic decision of describing HA→AA task delegation in terms of trust necessarily implies that the usual way of understanding human relations to technological products has become unsatisfactory. As many scholars note[27], relationships between hu-

[26] J. Buechner, H.T. Tavani, *op. cit.*, p. 43.

[27] M. Dzindolet *et al.*, *The role of trust in automation reliance*, in «International Journal of Human-Computer Studies», 58 (2003), pp. 697-718; P.J. Nickel *et al.*, *Can We Make Sense of the*

man users and technological products are usually framed in terms of *reliance*. Reliance might be said to indicate a property of relations to tools that refers directly to the function that a tool is supposed to carry out. Reliability, in turn, indicates the capacity of a tool to achieve the ends it is built to serve or, which is the same, «the ability of the item to remain functional»[28], thus forming, by being available, «the basis of new relations between its users and their environment»[29]. Accordingly, «we expect the artefact to function, to do what is meant to do as an instrument to attain goals set by humans»[30]. Deciding to frame HA→AA task delegation by reference to trust implies that, in such relations, something *more* is at stake that cannot be accounted for only by reference to the functional notion of reliance. What is this additional element?

Trying to answer this question is critical not only because of the reasons previously exposed, but also on the account that trust and reliance are not easily distinguishable. Although the importance of differentiating between trust and reliance is often stated, the word "trust" is just as often used as a synonym of "reliance". More precisely, the word "trust" appears to include the meaning of the word "rely" among its possible usages[31]. This is, beyond any doubt, what Taddeo means when she writes that we trust elevators. Similarly, Coeckelbergh[32] speaks of «'trust as reliance'»; Kiran and Verbeek, while discussing reliability, write that «tools can only be used for doing something if they are trustworthy»[33]; and Pitt defines untrustworthy technologies as «products that do not performed as promised, that break easily»[34]. In sum, as Nickel *et al.* observe, sometimes «what it means to

Notion of Trustworthy Technologies?, in «Knowledge, Technology and Policy», 23 (2010), pp. 429-444; K. Hawley, *Trust. A Very Short Introduction*, Oxford University Press, Oxford 2012, pp. 3-6.

[28] A. Birolini, *Reliability engineering. Theory and practice*, Springer, New York 2007, p. 2. Reliability is first of all a technical notion. However, once a technological product is deployed in social contexts, the technical measurement of its reliability is psychologically reinterpreted – sometimes in wrong ways. Distorted perceived reliability may lead to disuse due to underutilisation or misuse due to complacency; well perceived reliability leads to correct use and appropriate reliance and thus must be enforced (M. Dzindolet *et al.*, *op. cit.*; R.R. Hoffman *et al.*, *Trust in Automation*, in «IEEE Intelligent Systems», 28, 2013, n.1, pp. 84-88).

[29] A.H. Kiran, P.-P. Verbeek, *Trusting our Selves to Technology*, in «Knowledge, Technology, and Policy», 23 (2010), pp. 409-427, p. 422.

[30] M. Coeckelbergh, *op. cit.*, p. 54.

[31] P. Pettit, *Trust, Reliance and the Internet*, in «Analyse & Kritik», 26 (2004), pp. 108-121.

[32] M. Coeckelbergh, *op. cit.*, p. 54.

[33] A.H. Kiran, P.-P. Verbeek, *op. cit.*, p. 410.

[34] J.C. Pitt, *It's not about technology*, in «Knowledge, Technology and Policy», 23 (2010), pp. 445-454, pp. 450-451.

trust (to a certain degree) a technical artefact […] is more or less identical with what it means to rely (to a certain degree) on it»[35]. Sure enough, it would be of little significance to deny legitimacy to such usage. Nonetheless, the two words stand for different relationships that presuppose different scenarios and are based on different expectations, so that confusion on this point should be carefully avoided. It remains important, therefore, to reflect upon what is actually meant in these cases by the word "trust", and if something else is meant when the same word is applied in exclusively human contexts.

In order to clarify the role of trust in task delegation, let us consider HA→HA relations, where trust is commonly acknowledged as a constitutive element. When a person delegates a task to someone else, it is sensible to expect that she carries out an evaluation of the delegatee's overall adequacy to the task. Such adequacy is at least twofold: as Baier explains, «trust (…) is reliance on others' competence and willingness to look after, rather than harm, things one cares about which are entrusted to their care»[36]. The delegatee then must be *able* and *willing* to execute the appointed task[37]. Delegation will be successful if, and only if, the delegatee is both capable of carrying out the task appointed to her and inclined to commit to the delegator's requests. The first aspect concerns the delegatee's resources and skills, whilst the second pertains to her will or intent[38].

In delegation, reliability and trust emerge within these two dimensions. Reliability refers to the skills, abilities, and expertise that the delegatee possesses and exercises once delegation has occurred. A reliable person, regardless her trustworthiness, is competent, i.e., has what it takes to succeed in carrying out the task. In a sense, when someone's reliability is under scrutiny, she is already – even if partially – thought of as if she were a machine. She is indeed evaluated as an executer of predetermined tasks, i.e., of functions[39]. Reliability, as a measure of efficiency, is essentially relative to functional performances: when attributed to human beings, it

[35] P.J. Nickel *et al.*, *op. cit.*, p. 435.

[36] A. Baier, *Trust and Antitrust*, in «Ethics», 96 (1986), n. 2, pp. 231-260, p. 259.

[37] P. Pettit, *The cunning of trust*, in «Philosophy and Public Affairs», 24 (1995), n. 3, pp. 202-225; L.J. Camp *et al.*, *Trust: A Collision of Paradigms*, in P. Syverson (ed.), *Financial Cryptography*, FC 2001. Lecture Notes in Computer Science, vo. 2339, Springer, Berlin-Heidelberg 2002, pp. 91-105.

[38] A. Baier, *op. cit.*, p. 234.

[39] P.J. Nickel *et al.*, *op. cit.*, pp. 433-434; P. Pettit, *Trust, Reliance and the Internet*, cit., pp. 109-110.

indicates how well a person is able to serve as a means to a predetermined end in circumscribed contexts.

Trust, on the contrary, pertains to the second dimension of task delegation. Since human beings are free to choose what purposes to tend to, the delegator must secure the delegatee's commitment. Moreover, human beings can *fake* to tend to purposes other than those they actually tend to, so that assurance is even more required. Placing trust on the delegatee, who thus becomes a trustee, the delegator/trustor projects normative moral expectations on to the trustee's future behaviour[40]. In doing this, the trustor appeals to the trustee's sense of responsibility[41], impelling her to the task. As Luhmann writes, trust «serves to overcome an element of uncertainty in the behavior of other people which is experienced as the unpredictability of change in an object»[42]; or, in other words, its function is «the reduction of complexity in the face of the freedom of the other person»[43]. Trust puts social and moral pressure on the trustee, who is consequently motivated to align her own purposes to the trustor's ones. In the context of delegation, hence, trust is required when a delegatee, who is able to carry out the task, must also be persuaded to do so, since she may be uninterested in the delegator's demands or may fake interest, while having other purposes in mind. In this situation, trust adds an additional element to the relation between delegators and delegatees, which has the specific aim of motivating the latter to align their purposiveness to the former's one. Does the same need arise in HA→AA task delegation?

3. *Betrayal, Disappointment, and Robotrust*

Before addressing this question directly, it is necessary to spend few more words on why human beings place trust in order to delegate successfully. As already noted, trust enforces commitment, and commitment assures that human delegatees have assumed the delegators' end as their own. Since, when tasks are delegated, the delegatee's purposiveness may determine itself in counterproductive ways, measures must be taken so that it adjusts properly. In this context, trust is meant to influence the

[40] J. Buechner, H.T. Tavani, *op. cit.*, pp. 41-42.

[41] S.D.N. Cook, *Making the Technological Trustworthy: on Pitt on Technology and Trust*, in «Knowledge, Technology, and Policy», 23 (2010), pp. 455-459.

[42] N. Luhmann, *op. cit.*, p. 22.

[43] *Ivi*, p. 62.

delegatee's choice by placing a normative obligation on to her that appeals to her sense of responsibility. Feeling responsible not only for the task that has been appointed, but also for the trust that has been placed, the trustee is encouraged to carry out the delegated task and dissuaded to overwrite the trustor's purpose.

Consequently, in the context of HA→HA task delegation it is presumed that human beings enjoy a direct relationship to ends, i.e., that humans are self-determined purpose-setting agents. The delegator's attitude to trust stems from the presupposition that the delegatee has both the power to choose spontaneously among competing ends and preferences of her own. From such assumption follows that the delegatee may be uninterested in assuming the delegator's purpose or may fake to do so, while actually serving other purposes. The delegatee thus needs to be motivated so that she may truly take on ends set by somebody else. Trust, then, is a strategic response to nudge the unpredictable will of others by means of ethical and social pressure. Trust is required in HA→HA task delegation exclusively because it is assumed that HAs are *free* of determining their own will and choosing among different ends.

When a human delegatee, who acts as if she has taken on a given task, does not truly commit to the trustor's aim or has a personal agenda that collides with it, a breach of trust occurs. If a trustee fails to carry out the task entrusted to her, even though she was perfectly capable of executing it, the trustor feels *betrayed*[44]. In task delegation, betrayal is the accusation that trustors throw to noncomplying trustees, that is, to trustees who determine their own will regardless of their commitment to the delegators' purpose. Since normative moral expectations were placed on the trustee's behaviour, the trustor has the right to complain, to hold the delegatee morally responsible for breaching trust, to ask for justification or explanation, and perhaps even to take offence. In sum, the notion of trust in the context of HA→HA task delegation describes a property of such relation that has the function of minimising betrayal. This result, in turn, is accomplished through a dialectic of normative moral expectation and reactive responsible behaviour that nudges the delegatee's will in a way that fosters successful delegation.

In light of this, asking whether trust is needed in HA→AA task delegation equals asking whether AAs are capable of choosing autonomously among different ends and, thus, whether they must be motivated accord-

[44] A. Baier, *op. cit.*, p. 235; H. Tavani, *op. cit.*, pp. 86-88.

ingly so as to align their preferences to the delegator's ones. In other words, what must be clarified is whether AAs are self-determined purpose-setting entities, since only such entities can betray, thus making trust necessary. If yes, trust is required in HA→AA task delegation just as it is required in HA→HA task delegation, and it is correct to transfer trust from human to artificial delegatees. If not, AAs would be entities that serve predetermined ends. Thus, the only dimension of task delegation that would remain would be that of ability or efficiency. In this case, reliance should suffice to understand HA→AA task delegation and the introduction of trust would be spurious.

In my opinion, it is not possible to think AAs as entities which can spontaneously choose among competing ends and betray, since they do not exhibit a direct relation to purposes. No AA «*has* purpose and *acts on* purpose»[45] the way we do. While human beings are purpose-setting entities – or at least are supposed to be so in the context of task delegation – AAs are «purpose-built artifacts»[46] as any other technological product and there is no need to assume them to be anything more. Even though AAs display the distinguishing feature of being able of executing functions independently from human oversight or intervention, they are still fully understandable by reference to the category of tool[47]. As any other tool, AAs require a specified end to serve in order to be devised, designed, and manufactured. It is impossible to think AAs apart from the specific purposes they are built to serve – which are, therefore, always *given*.

Although AAs can carry out functions autonomously and, consequently, partially unpredictably (as no previous tool could), still they do not display the possibility of setting ends by themselves nor of intentionally serving unpredictable ends. Accordingly, tasks are delegated to AAs only in virtue of their efficiency in achieving ends that are valuable for their users. The range of AAs' autonomy and unpredictability extends exclusively to the execution of functions, that is, to the way in which given ends are accomplished. AAs serve a purpose or clusters of purposes that can always be traced back to their designers. At the same time, this purpose or these clusters of purposes identify with the reasons why they appear useful. AAs

[45] H. Jonas, *The Phenomenon of Life. Towards a Philosophical Biology*, Northwestern University Press, Evanston 2001, p. 119.

[46] J.J. Bryson, P. Kime, *Just an Artifact: Why Machines are Perceived as Moral Agents*, https://www.cs.bath.ac.uk/~jjb/ftp/BrysonKime-IJCAI11.pdf [accessed 1 October 2018], p. 1.

[47] F. Fossa, *Artificial Moral Agents: Moral Mentors or Sensible Tools?*, in «Ethics and Information Technology», 20 (2018), pp. 115-126.

are purpose-built artefacts that exist only within socio-technical contexts where ends are set and pursued[48].

As it seems sensible to frame AAs as purpose-built artefact, it seems also sensible to deny the possibility that AAs can betray[49]. Lacking the capacity of setting ends autonomously, AAs can neither be uninterested in the task they execute nor fake interest in the task, while intentionally serving other purposes in secret[50]. The alignment of any AA to the end of the function it executes is in fact merely a matter of design. Well-designed AAs will carry out their task as they are supposed to, whilst poor-designed AAs will not. In the case of HA→AA task delegation, then, there are no conditions for trust to emerge. Accordingly, true betrayal cannot occur in this context, but can be experienced only metaphorically.

When an AA fails to achieve the goal it is programmed to pursue, users ought not to interpret this failure in terms of betrayal, but rather in terms of *disappointment*[51]. Disappointment refers to functional expectations that are not met and, as such, is the appropriate reaction to reliability issues. AAs that repeatedly disappoint their users are unreliable – and "untrustworthy" only in this technical sense. Accordingly, it would be irrational to take offence at AAs or holding them morally responsible for failing to fulfil their commitment[52], just as it would be irrational to take offence at an elevator or hold it morally responsible in case of incident. Unreliable AAs can either be discarded or fixed so that they carry out their function as originally intended, in the most effective way possible. There are no untrustworthy AAs, just malfunctioning or poorly designed ones. In HA→AA task delegation, only the dimension of reliability emerges.

For this reason, in the case of task delegation there is no actual need to

[48] D. Johnson, *Computer Systems. Moral Entities, but Not Moral Agents*, in «Ethics and Information Technology», 8 (2006), n. 4, pp. 168-183.

[49] J. Simon, *op. cit.*, pp. 346-347; A. Van Wynsberghe, S. Robbins, *Critiquing the Reasons for Making Artificial Moral Agents*, in «Science and Engineering Ethics» (2018), https://doi.org/10.1007/s11948-018-0030-8 [accessed 1 October 2018].

[50] This does not mean, of course, that no technological product may run some functions explicitly while at the same time running other functions implicitly which do not align with the user's intentions. When this happens, however, it is not due to a lack of motivation in the AA, but it happens *by design*. Such AA would still be reliable, since it would do what is supposed to; however, the *social description* of the AAs would be untrustworthy, since it would not inform the users on what the AA does. This problem will be discussed in Section 5.

[51] L.J. Camp *et al.*, *op. cit.*; J.C. Pitt, *op. cit.*

[52] P. de Laat, *Trusting the (Ro)botic Other: By Assumption?*, in «SIGCAS Computer and Society», 45 (2015), n. 3, pp. 255-260.

move from reliance to trust once AAs are involved, even though they are, in a sense, autonomous and unpredictable entities. Framing HA→AA task delegation in terms of reliance rather than in terms of trust is not only more accurate, but also safer since it prevents anthropomorphism and the many social risks deriving from it[53]. Since trust is essentially linked to the need of steering someone else's will through motivation, placing trust in AAs would presuppose the existence of a practical dimension, that of purpose-setting freedom, which does not belong to artificial agency. On the contrary, such practical dimension characterises human agency, so that projecting purpose-setting freedom onto AAs would result in humanising them. Consequently, the relation with AAs might appear to require measures and precautions that, yet, would be misplaced. Moreover, possible malfunctions would risk being mistakenly charged with moral meaning while, at the same time, ill-suited moral expectations would develop. As Bryson notes, anthropomorphising AAs «invites inappropriate decision such as misassignations of responsibility and misappropriations of resources»[54]: framing HA→AA task delegation in terms of trust would probably risk a similar social effect. Finally, understanding AAs as direct objects of trust might divert attention from the social context in which HA→AA task delegation occurs – i.e., from the context related to AAs where trust does play a crucial role.

4. *Artificial Agents and Social Trust Mediation*

In light of what has been said, it proves more accurate to interpret direct HA→AA relations by reference to the notion of reliance, rather than to the notion of trust. Strictly speaking, trust does not pertain to the relationship between users and artefacts, though advanced they may be. To some extent, however, confusion on this matter may arise since trust permeates the social context in which HA→AA task delegation occurs. While executing delegated tasks, in fact, AAs mediate not only human agency, thus saving time and resources, but also trust relationships between various social actors. Being always embedded in social contexts, AAs become inter-

[53] P.J. Nickel *et al.*, *op. cit.*

[54] J.J. Bryson, *Robots Should Be Slaves*, in Y. Wilks (ed.), *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues*, John Benjamins Publish Company, Amsterdam 2010, pp. 63-74, p. 64.

section points of ethical normative expectations and responsibilities. In this *indirect* sense – i.e., as social trust mediators – AAs are the cornerstone on which trust relations between social actors are built. Such relations, moreover, play a critical role in shaping the social attitude towards automation, so that it is crucial neither to overlook such dimension nor to let it fade behind the misconception of direct HA→AA relationship as involving trust. Trust should not be mistakenly extended from the social milieu to the HA→AA relation itself: the two levels must not be confused.

In this light, a discussion concerning social trust as mediated by AAs is both possible and extremely relevant. However, such discussion can be properly carried out provided that, in the study of direct HA→AA relationships, trust is set aside. Only once it has been clarified why AAs cannot be trustees it becomes possible to ask who is truly trusted, when tasks are delegated to AAs.

In order to clarify in what sense AAs mediate social trust it is necessary to take a closer look to HA→AA task delegation. When a person considers whether to delegate tasks to an AA, it is rational to expect that she would try to establish if the AA in question "will do", that is, if it is capable of executing the desired task. Such evaluation concerns the AA's efficiency: it ponders over what purpose the AA is supposed to serve and how effective it is supposed to function. However, how can one know what a particular AA is supposed to do? Either the delegator has a deep understanding of the technology involved – which is arguably a rare case – or she will have to turn to a nontechnical description of the AA, to which she can meaningfully relate[55]. It follows that delegation will be most likely grounded on a description of the AA provided by those who happen to have the necessary expertise to express the AA's specifics in common language. Therefore, trusting the product means trusting the nontechnical description of its functions or utility; and this, in turn, means trusting the social actors who provide such description.

In HA→AA task delegation, then, direct relations between users and technologies are embedded in indirect relations between users and those who provide nontechnical descriptions of AAs. For the sake of the present argument, I will address those social actors as "stakeholders"[56]. Trust

[55] W. Pieters, *Explanation and trust: what to tell the users in security and AI?*, in «Ethics and Information Technology», 13 (2011), pp. 53-64.

[56] With the label "stakeholders" I mean all the subjects who are in different degrees involved in providing end-users with a comprehensible description of technological objects. In this specific sense, the label may apply to designers, programmers, engineers, advertisers, firms,

relations between end-users and stakeholders revolve around the adequateness of the AA's nontechnical description[57]. If this description is satisfying, the users will not need to worry about anything else than the AA's reliability. However, the description may be flawed or biased. For example, it may turn out that the AA executes other functions than those described, that it employs other means than those indicated, or that it also carries out undisclosed tasks. If any of these (or other) cases occur, users may reasonably feel betrayed; and since betrayal is a marker of trust, it suggests that the relation between end-users and stakeholders is one of trust. Inadequate descriptions lead to breaches of trust and, to this extent, causes AAs to appear untrustworthy (even if, perhaps, reliable).

The relation between users and stakeholders may be characterised as a case of HA→HA task delegation. HA→AA task delegation always occurs within a social context where AAs are presented to the public, their utility and features are advertised, and delegation itself is often encouraged. Knowingly or unknowingly, users delegate to stakeholders the task of providing an adequate description of the product they offer. This task, in fact, cannot be performed directly by the end-users, since they usually lack the necessary knowledge; and even if they could, to analyse meticulously every device one would want to use would still be extremely demanding in terms of time and resources. When AAs pass from the stakeholders' on to the end-users' hands, they bring along a description of the functions they carry out and the ends they serve, which translates the specifics of the artefacts in a user-friendly language. This description is the result of a human activity; and human beings can fake interest in delegated tasks, while intentionally serving other ends. Therefore, in this practical situation the conditions for trust apply. Users (as trustors) entrust to stakeholders (who become trustees) the task of providing a nontechnical description of the AA that would not be biased, incomplete, malevolent, or opaque. Placing

companies, institutions and so on. However, also the work of science communicators, journalists, and artists deeply influences the social understanding of technological products, which is then constantly negotiated. In its present form, the category is evidently ill-defined; nonetheless, it meets the need for which it is introduced in the argument. Further clarifications must be postponed. A similar use of the term "stakeholders" may be found in M. Hengstler *et al.*, *Applied Artificial Intelligence and Trust*, in «Technological Forecasting and Social Change», 105 (2016), pp. 105-120; P. de Laat, *op. cit.*; D. Pedreschi *et al.*, *Open the Black Box. Data-Driven Explanation of Black Box Decision Systems*, https://doi.org/0000001.0000001

[57] What elements make a nontechnical description "adequate" (transparency, honesty, completeness, clarity and so on) is an issue that requires much discussion and cannot be dealt with here.

trust, in this peculiar case, has the purpose of minimising the chance of betrayal by means of social and moral pressure. Hence, AAs indirectly mediate trust relationships between different social actors, i.e., *social trust*. "Untrustworthy" technologies are not such in themselves, but as devices made by untrustworthy producers or deployed by untrustworthy subjects.

The dimension of indirect trust mediation in HA→AA task delegation must not be overlooked, since much of the social acceptance of AAs depends from it[58]. Whether users will consider stakeholders trustworthy or not will affect their general disposition towards social robotics and AI systems in important ways. Low trust in stakeholders impinges significantly on the success of task delegation to AAs. Well-designed, reliable technologies will always appear in a suspicious light unless the companies and institutions behind them make an effort to earn the users' trust.

The trustworthiness of those who provide nontechnical descriptions of AAs represents undoubtedly a relevant issue to be addressed in the future from a social viewpoint. On this account, both ethical reflection and legal regulation must take on the task of indicating, recommending, and enforcing the right means to protect and maximise social trust. In conclusion, trusting AAs beyond reliance means trusting their nontechnical description and, thus, the social actors who provide it. Exclusively in this indirect, social sense, it seems possible to discuss trust, breach of trust, and distrust in situations involving AAs – which is entirely different from understanding AAs directly as trustees. Facilitating, protecting and enhancing trust between the human beings whose actions are practically mediated by AAs may be one of the most critical challenges posed by AI and robotics to future society.

Abstract

The aim of this paper is to clarify the extent to which relationships between Human Agents (HAs) and Artificial Agents (AAs) can be adequately defined in terms of trust. Since such relationships consist mostly in the allocation of tasks to technological products, particular attention is paid to the notion of delegation. In short, I argue that it would be more accurate to describe direct *relationships between HAs and AAs in terms of reliance, rather*

[58] M. Hengstler *et al.*, *op. cit.*; A.F. Winfield, M. Jirotka, *Ethical governance is essential to building trust in robotics and AI systems*, in «Philosophical Transactions A: Mathematical, Physical and Engineering Sciences», 2018 [in press].

than in terms of trust. However, as mediums of human actions to which tasks are delegated, AAs indirectly mediate trust between users and other social actors involved in their design, manufacture, commercialisation and deployment. In this sense, AAs mediate social trust. My conclusion is that relationships between HAs and AAs are thus to be understood directly in terms of reliance and indirectly in terms of social trust mediation.

Keywords: artificial agents; trust; robotrust; reliance; human-robot interaction.

Fabio Fossa
Università di Pisa
fabiofossa36@gmail.com

T

Placing Trust in Medicine by Dealing with Its Uncertainty

Francesca Marin

1. *The need for a triple pattern*

This paper promotes the idea of mutual dependence between trust, medicine and uncertainty, and critically analyzes those approaches that either partially recognize such interdependence or propose a misrepresented view of medicine in terms of its relationship with uncertainty, negatively affecting trust in medicine. Indeed, on the one hand, nowadays the trust-medicine dyad is sometimes recognized without the acknowledgment of the medicine-uncertainty dyad, or vice versa. In other words, the role of trust in medical practice and the presence of uncertainty in medicine could be approached as issues which are unrelated to each other. In this way, the process of planning to promote trust in medicine and strategies for responding to medical uncertainty might be separately addressed. On the other hand, intolerance or even a refusal of medical uncertainty could affect trust in medicine because, by considering medicine itself as a science and a practice characterized by full certainty, claims of infallibility on the one side, and suspicions as well as incredulity on the other, might be fostered.

In order to avoid all these reductive views and to promote a well-placed trust in medicine, the paper's aim is to argue for the need for a change from a dual scheme, i.e. trust-medicine or medicine-uncertainty, to a triple pattern, that is, the trust-medicine-uncertainty interdependency. This is a particularly innovative proposal because, as it will be argued in the next paragraph, the scientific literature does not seem to have examined the trust-medicine-uncertainty pattern in depth. Indeed, although there are some exceptions[1],

[1] For example, see K. Armstrong, *If You Can't Beat It, Join It: Uncertainty and Trust in*

 DOI: 10.4454/teoria.v39i1.58

the debate has been more concerned with analyzing the twofold dyad mentioned above, i.e. trust-medicine and uncertainty-medicine. In this way, the role of trust in the medical context and the implications of uncertainty for both the clinical encounter and healthcare systems have been addressed. Although relevant, these contributions to the debate seem to have failed in acknowledging the trust-medicine-uncertainty interdependency, leading one to consider any effort to hide or eliminate medical uncertainty as a particularly promising strategy to promote trust in medicine.

The first part of the paper aims to critically analyze this strategy, not only by asserting that medical uncertainty cannot be completely removed, but also by arguing that there is a kind of irreducible uncertainty that typifies the nature of medicine as a science and a practice. This is an intrinsic uncertainty due to the epistemological status of medicine that cannot be confused with other forms of medical uncertainties, such as those arising from personal factors or limits of available medical knowledge. This is why the main forms of medical uncertainty will firstly be distinguished and facing one of them, intrinsic uncertainty, will secondly turn out to be a necessary condition for *well-placed trust in medicine*. This approach will be suggested by exploring the main consequences arising from the unwillingness to acknowledge and tolerate intrinsic medical uncertainty. In particular, examples of *misplaced trust in medicine* due to considering medicine as an absolutely certain scientific knowledge[2] and *misplaced distrust in medicine* as a result of an antiscientific view of medical knowledge will be discussed.

In the second part of the paper, the need to promote trust in medicine by dealing with its uncertainty will prove to be particularly urgent. Strictly speaking, the words "evidence" and "precision", which are abundantly used in the era of evidence-based medicine and precision medicine, could erroneously suggest a high degree of certainty and thus obscure the irreducible

Medicine, in «Annals of Internal Medicine», 168 (2018), n. 11, pp. 818-819; J.B. Imber, *How Navigating Uncertainty Motivates Trust in Medicine*, in «AMA Journal of Ethics», 19 (2017), n. 4, pp. 391-398.

[2] This view is usually based on the so-called "scientism", which rests on a problematic epistemological tenet. According to this tenet, only science can provide a successful explanation of the reality so much so that «scientific inquiry is our only genuine source of knowledge; all other alleged forms of knowledge (e.g., ordinary perception, *a priori* knowledge and introspection) are either reducible in principle to scientific knowledge or illegitimate». M. De Caro, *Realism, Common Sense, and Science*, in «The Monist», 98 (2015), pp. 197-214, p. 203. For an in-depth critical analysis of scientism and for a more inclusive approach to nature than any provided by the natural sciences, see M. De Caro, D. Macarthur (eds), *Naturalism in Question*, Harvard University Press, Cambridge (MA) 2004.

uncertainty of medicine. Actually, advances in these fields, rather than diminishing medical uncertainty, are contributing to its increase and even generating new kinds of uncertainties. After providing some examples about this topic, it will be argued that adequately promoting the expansion of medical knowledge means further justifying the trust-medicine-uncertainty interdependency. In other words, it means acknowledging that the triple pattern proposed in this paper is advantageous, and no less than necessary.

2. *A twofold dyad: trust-medicine and uncertainty-medicine*

The crucial role of trust in medical practice has long been recognized, and reasons to encourage mutually trusting relationships in medical context have been addressed. For instance, the phenomenological approach to the clinical encounter proposed by Edmund Pellegrino stresses that being ill means experiencing a particular vulnerable state as well as being forced to seek assistance from and to trust another person, i.e. the health professional, who holds the balance of the power by having the necessary knowledge and competences to heal[3]. As a consequence, strategies for building trust are both requisite responses to the patient's vulnerability and efforts to reduce the inequality of knowledge and skills that characterizes any relation with professionals, and consequently the patient-physician relationship as well[4].

Besides being required by the nature of the clinical encounter, interpersonal trust is also rightly considered one of the most important contributors to effective care. In medical settings, trusting attitudes such as loyalty, willingness to listen, truthful communication, and empathy usually result in improved health outcomes. Actually, good health care requires trust in medical institutions and health care systems as well, so much so that organizational aspects, such as availability of and accessibility to health care services, are likely to affect trust in one's doctor and, conversely, a trusting patient-physician relationship may enhance institutional trust[5].

[3] E.D. Pellegrino, *Toward a Reconstruction of Medical Morality: The Primacy of the Act of Profession and the Fact of Illness*, in R. Bulger, J. McGovern (eds), *Physician and Philosopher. The Philosophical Foundation of Medicine: Essays by Dr. Edmund Pellegrino*, Carden Jennings Publishing, Charlottesville 2001, pp. 18-36.

[4] Cfr. C.C. Clark, *Trust in Medicine*, in «Journal of Medicine and Philosophy», 27 (2002), n. 1, pp. 11-29.

[5] J. Saunders, *Trust and Mistrust Between Patients and Doctors*, in T. Schramme, S. Edwards (eds), *Handbook of the Philosophy of Medicine*, Springer, Dordrecht 2017, pp. 487-502.

These considerations explain why strengthening trust in medical science and practice is a primary goal. Nevertheless, in order to achieve such a purpose, it seems necessary to come to terms with another feature of medicine, its uncertainty, which has long been addressed within scientific literature. Indeed, both the philosophy of medicine and medical sociology have analyzed the medicine-uncertainty dyad, examining sources and implications of medical uncertainty for the clinical encounter[6] as well as for healthcare systems[7]. Uncertainty related to, for example, diagnosis and outcome is generally uncomfortable for those involved in the clinical decision-making process. Indeed, such uncertainty might instill further vulnerability in the patient and evoke a sense of helplessness in the physician. As a consequence, questions have been raised about whether and how uncertainty should be communicated[8] as well as whether and how physicians should be trained for uncertainty[9]. Furthermore, inappropriate responses to uncertainty might have a negative impact on the quality and cost-effectiveness of healthcare system, for example leading to unnecessary diagnostic tests or treatments.

[6] M.S. Henry, *Uncertainty, Responsibility, and the Evolution of the Physician/Patient Relationship*, in «Journal of Medical Ethics», 32 (2006), pp. 321-323.

[7] R.L. Logan, P.J. Scott, *Uncertainty in Clinical Practice: Implications for Quality and Costs of Health Care*, in «Lancet», 347 (1996), pp. 595-598.

[8] P.K. Han, *Conceptual, Methodological, and Ethical Problems in Communicating Uncertainty in Clinical Evidence*, in «Medical Care Research and Review», 70 (2013), 1 Suppl., pp. 14-36. Many qualitative studies have investigated this topic, obtaining non-homogenous results about the disclosure of uncertainty by physicians and patient satisfaction deriving from such disclosure. For example, Braddock and colleagues reported that during the informed consent process, physicians shared uncertainty with their patients in only 5% of the clinical encounters. See C.H. Braddock, K.A. Edwards, N.M. Hasenberg *et al.*, *Informed Decision Making in Outpatient Practice: Time to Get Back to Basics*, in «JAMA», 282 (1999), pp. 2313-2320. In contrast, in a more recent audiotape study, physicians made verbal expressions of uncertainty in 71% of clinic visits. Cr. G.H. Gordon, S.K. Joos, J. Byrne, *Physician Expressions of Uncertainty During Patient Encounters*, in «Patient Education and Counseling», 40 (2000), pp. 59-65. Furthermore, in the study conducted by Gordon, an increase has been registered in patient satisfaction with physician expression of uncertainty. Nevertheless, another study focused on therapeutic uncertainty has shown that patient satisfaction ratings were highest when no uncertainty was shared by the physician. See C.G. Johnson, J.C. Levenkron, A.L. Suchman *et al.*, *Does Physician Uncertainty Affect Patient Satisfaction?*, in «Journal of General Internal Medicine», 3 (1988), pp. 144-149. The results of the same study have confirmed that patient satisfaction is influenced by the manner in which uncertainty is conveyed and resolved by the physician.

[9] For opposite views on this issue, see R.C. Fox, *Training for Uncertainty*, in R.K. Merton, G. Reader, P.L. Kendall (eds), *The Student-Physician. Introductory Studies in the Sociology of Medical Education*, Harvard University Press, Cambridge 1957, pp. 207-241 and P. Atkinson, *Training for Certainty*, in «Social Science & Medicine», 19 (1984), n. 9, pp. 949-956.

All these considerations might suggest the idea that efforts to hide or remove medical uncertainty are a particularly promising strategy to promote trust in medicine. As it will be argued later, in addition to being unattainable given that medical uncertainty cannot be completely eliminated, such a project is disadvantageous because it deprives medicine of its proper nature and could lead to the decline of the authoritativeness of medical knowledge.

3. *The main kinds of uncertainty in medicine*

Quoting David Eddy, «uncertainty creeps into medical practice through every pore» so much so that the patient-physician relationship could be described as "a chain of uncertainty"[10]. From achieving more knowledge about a patient's condition to selecting and following a treatment plan, there are several links of uncertainty that vary depending on who performs the procedure and upon whom it is performed. These are what Eric Beresford has defined as "personal sources of uncertainty[11] alluding to both patient factors and physician aspects. Among the former, for example, there are biological variability, variable responses to treatment, partial presentation of symptoms to the physician, access to other sources of information, and even incompetence or inability to make wishes known. Diagnostic, prognostic and therapeutic uncertainty might be further emphasized by physician factors, such as bias, personal ignorance, poor communication skills, and intolerance to acknowledge the actual limits of medical information[12]. Furthermore, medical practice is characterized both by what Beresford has called "conceptual uncertainty", which occurs when applying abstract criteria (i.e. treatment guidelines or risk classifications) to particular patients, and by «uncertainty arising from health-care management and delivery, related to the complexity of systems involving a myriad of healthcare professionals that need to be coordinated, managed, and regulated»[13].

[10] D.M. Eddy, *Variations in Physician Practice: The Role of Uncertainty*, in «Health Affairs», 3 (1984), pp. 74-89, p. 75.

[11] E.B. Beresford, *Uncertainty and the Shaping of Medical Decisions*, in «Hastings Center Report», 21 (1991), n. 4, pp. 6-11.

[12] T. Dhawale, L.M. Steuten, H.J. Deeg, *Uncertainty of Physicians and Patients in Medical Decision Making*, in «Biology of Blood and Marrow Transplantation», 23 (2017), pp. 865-869 (in particular, p. 867).

[13] A.J.E. Seely, *Embracing the Certainty of Uncertainty. Implications for Health Care and Research*, in «Perspectives in Biology and Medicine», 56 (2013), n. 1, p. 68. For further types of

In addition to personal and conceptual uncertainties as well as to those originated by healthcare systems, there is a kind of uncertainty due to limitations in currently available medical knowledge and thus based on a scientific data deficit. This is the so-called "technical or informational uncertainty", which is expected to be reduced by the continuous progress in medical research. Nevertheless, the legitimate goal of acquiring additional knowledge cannot be rooted in the belief that this acquisition will totally remove medical uncertainty. Such a belief is unfounded not only because medical practice is confronted with the complexity and the singularity of the particular on a daily basis, but also because medicine is marked by an "irreducible or intrinsic uncertainty". Being a science and a practice, medicine discloses an alterable character given that medical knowledge and competences are all, at least in principle, revisable and no physician's mind is a blank slate or *tabula rasa*, but rather "a mind of a physician". Quoting Dario Antiseri, «behind a physician's eyes and hands there is a *mind of a doctor* and this *mind of a doctor* is laden with theories, expectations, experiences, mistakes already made by himself and by other physicians, technical devices, therapeutic theories, solved (and unsolved) clinical cases»[14]. The previous points are, as already mentioned, open to revision. As a consequence, intrinsic uncertainty is an essential characteristic of medicine and is an inevitable companion of medical practice.

It must be noted that the acknowledgment of this hallmark of medicine is not in conflict with any effort to decrease informational uncertainty. Not only is intrinsic uncertainty different from informational uncertainty, but also accepting the former as an inherent feature of medicine is a necessary condition for fostering our desire to reduce the deficit in current medical knowledge[15] as well as for being willing to minimize medical uncertainty. In order to examine this point in depth, some epistemological considerations regarding scientific knowledge are required. Although the following notes might initially appear misplaced, they will later reveal themselves to be an argument that justifies the need for the triple pattern proposed in this paper.

medical uncertainty, see P.K.J. Han, W.M.P. Klein, N.K. Arora, *Varieties of Uncertainty in Health Care: A Conceptual Taxonomy*, in «Medical Decision Making», 31 (2011), n. 6, pp. 828-838.

[14] D. Antiseri, *Epistemologia contemporanea e logica della diagnosi clinica*, in P. Giaretta, A. Moretto, G.F. Gensini, M. Trabucchi (eds), *Filosofia della medicina. Metodo, modelli, cura ed errori*, il Mulino, Bologna 2009, pp. 75-104, p. 81 (my translation).

[15] Although softer, a similar argument has been proposed by Seely when stating that «accepting intrinsic uncertainty is complementary to our desire to reduce and quantify informational uncertainty». A.J.E. Seely, *art. cit.*, p. 67.

4. *Toward a well-placed trust in medicine*

By its nature, scientific enterprise is revisable and historical evidence confirms this aspect so much so that scientific theories are likely to be, in Popper's terms, falsified or, better, superseded by other scientific theories. As a consequence, the acquisition of further knowledge in any scientific field must be supported by the acknowledgement that no area of scientific knowledge is characterized by a degree of absolute or apodictic certainty. In fact, any assertion or theory that would present itself as a totally irrefutable or incontrovertible knowledge could not be considered as a scientific statement[16]. The revisable character of scientific enterprise unavoidably assigns to any content of science a certain degree of uncertainty. Due to the particular nature of science, this is an uncertainty that, although modifiable, is ineradicable.

It is important to note that the certainty of such uncertainty does not question the authoritativeness of scientific knowledge. On the contrary, it is properly the source of this uncertainty, that is, the revisability of scientific knowledge, which makes science an authoritative form of knowledge. Indeed, such revisability characterizes the slow "march of science"[17], but does not invalidate the truth of scientific knowledge: although scientific truth is temporary precisely because it is revisable, it is however a scientific truth by being evidence-based as well as partially or entirely accredited by the scientific community. Furthermore, the revisable character of scientific enterprise can firstly guarantee an even more in-depth knowledge of reality, and secondly, impede that any scientific theory exclusively auto-confirms itself.

Applying these considerations to medicine, we can say that what is often considered as a failure of medicine or its Achilles' heel, that is, its irreducible uncertainty and thus its continual process of revision, assigns to medical knowledge the status of scientific knowledge. In other words, it is precisely this fickle aspect which is the greatest strength of medicine because it makes medicine an authoritative form of knowledge. This explains why acknowledging the medical intrinsic uncertainty means recognizing the epistemological status of medicine and differentiating such uncertainty from personal, conceptual and informational uncertainties is the starting point for dealing with medical uncertainty and being willing to reduce it.

[16] Cfr. E. Agazzi, *Scientific Objectivity and Its Context*, Springer, London 2014, p. 411.

[17] L. Rosenbaum, *The March of Science - The True Story*, in «The New England Journal of Medicine», 377 (2017), n. 2, pp. 188-191.

Indeed, any effort in this direction should consider a complete removal of intrinsic uncertainty to be not only impossible, but also disadvantageous or even undesirable because it would lead to the decline of medicine and of the authoritativeness of medical knowledge.

Accepting this conclusion means adopting an approach for the promotion of trust in medicine that is not aimed to deny or hide intrinsic uncertainty, but rather to acknowledge and value it. To specify, this is an approach that encourages us to put trust in medicine by precisely facing and embracing its intrinsic uncertainty. Such encouragement is rooted in the trust-medicine-uncertainty interdependence because it does not simply address the twofold dyad, that is, the crucial role of trust in medical context and the ubiquitous presence of uncertainty in medical practice, but also considers uncertainty as a fundamental aspect of the epistemological status of medicine, whose presence guarantees well-placed trust in medicine.

5. *Misplaced trust vs misplaced distrust in medicine*

By analyzing medical uncertainty and in particular the kind of uncertainty that intrinsically characterizes medicine as a science and a practice, the link between the revisability of scientific knowledge and the authoritativeness of science has been addressed. Nowadays this link is not easily recognized and accepted, so much so that a sort of intolerance toward the revisability of medical knowledge is still widespread not only in the public opinion, but also within a part of the scientific community. This intolerant attitude could assign the highest degree of certainty and absoluteness to medical knowledge, misunderstanding the proper nature of medicine and leading to unquestioning trust in medicine.

Besides offering an epistemologically problematic scenario, this misplaced trust in medicine affects medical practice and raises many ethical issues. Firstly, if medical knowledge is considered as the only knowledge deserving of the name "scientific knowledge", a standardized approach to disease could be proposed, reducing the human body to a completely quantifiable reality as well as determining any status of health exclusively on the basis of objective parameters. In this way, our body might be considered as a mere extended physical substance (*Körper*), underestimating the subjective experience of our corporeity (*Leib*)[18], and personal valua-

[18] In this respect, see E. Dahl, C. Falke, T.E. Eriksen (eds), *Phenomenology of the Broken*

tions regarding our health status could be totally excluded. Secondly, efforts to consider or even to present medical knowledge as an absolutely certain form of knowledge usually result in a claim for infallibility and in a search for the highest degree of certainty. To specify, on the one hand, denying or hiding medical intrinsic uncertainty could increase the degree of patient expectations and create new demands towards medicine. On the other hand, an unwillingness or incapacity to accept intrinsic uncertainty could lead to a physician's maladaptive responses to uncertainty, such as anxiety, obsession with finding the right answer, and reluctance to disclose uncertainty for fear of projecting ignorance or failure to patients[19].

Further problematic issues arise when the revisability of scientific knowledge is recognized but the authoritativeness of science is questioned or, in the worst case, denied. When dismissing scientific consensus, perceptions of corruption are usually invoked. Unfortunately, stories of scientific misreporting (such as exaggeration of the conclusions drawn from research and unpublished negative findings)[20], conflicts of interest with pharmaceutical industry and political meddling are, although rare, sadly true. Nevertheless, doubts, suspicions and skepticism are often unwarranted, and an antiscientific view of medical knowledge is also due to an intolerance of medical uncertainty. Indeed, it is precisely this fickle aspect of medicine that generally leads people to distrust medical data and guidelines, for example by resorting to alternative medicines. In this way, the authoritativeness of medical knowledge is usually superseded by emotions, personal beliefs and pseudoscientific conspiracies whose misjudged statements are easily shared and exponentially amplified by social platforms. It's no coincidence that the Oxford Dictionaries declared "post-truth" as the international Word of the Year for 2016, defining it as an adjective «relating to or denoting circumstances in which objective facts are less influential in shaping public opinion than appeals to emotion and personal belief»[21].

Body, Routledge, London 2018; R.T. Jensen, D. Moran (eds), *The Phenomenology of Embodied Subjectivity*, Dordrecht, Springer 2013; S. Gallagher, D. Zahavi, *The Phenomenological Mind. An Introduction to Philosophy of Mind and Cognitive Science*, Routledge, London 2012[2].

[19] A.L. Simpkin, R.M. Schwartzstein, *Tolerating Uncertainty - The Next Medical Revolution?*, in «The New England Journal of Medicine», 375 (2016), n. 18, pp. 1713-1715.

[20] For some misleading conclusions drawn from health-related research, see P. Sumner, S. Vivian-Griffiths, J. Boivin *et al.*, *The Association Between Exaggeration in Health Related Science News and Academic Press Releases: retrospective Observation Study*, in «British Medical Journal», 349 (2014), https://www.bmj.com/content/bmj/349/bmj.g7015.full.pdf.

[21] https://en.oxforddictionaries.com/word-of-the-year/word-of-the-year-2016.

Passively accepting to live in a post-truth era means dangerously impeding the spread of evidence-based data and calling into question the strength of rational arguments. When dealing with health-related issues, this tendency is potentially harmful both for the individual and for society because in general fake news, fallacies, inconsistent beliefs and scams about unsubstantiated treatments spread useless, and even worse, dangerous practices, usually for commercial purposes or for making money easily. This risk is further exacerbated by the advancement of Information and Communication Technologies (ICT). Indeed, through the use of search engines, users may be subjected to the so-called "information bubble phenomena", that is, being isolated in a universe of information algorithmically created on the basis of users' location, personal preferences and past click-behavior[22].

The weakest people, such as teenagers and the elderly, as well as those who uncritically use search engines and social networks, are more likely to be exposed to misinformation in the medical field and prone to the information bubble. Their searching on the internet could be exploited by ICT companies, for example for displaying advertisements of selected goods or unsubstantiated therapies during users' future online browsing, which might appear to vulnerable people as "the solution" to their problems.

To sum up, living in a digital and post-truth era where it is becoming increasingly necessary to verify truthfulness and the quality of information, it is more reasonable to sustain the slow "march of science" and accept the revisability of scientific knowledge than to be damaged by unverified or false information. In fact, as sadly confirmed by the news, a believed lie (for example believing that cancer might be cured with sodium bicarbonate) can cause irreversible injury, and, in the worst-case scenario, lead to death. As a consequence, questioning medical knowledge although there is a reasonable evidence for trustworthiness, and at the same time unquestioning non-authoritative opinions as well as untrustworthiness information obtained in the internet or shared by social platforms, is a real paradox. Even more, it is proof of an irresponsible behavior that leads to misplaced distrust in medicine and has negative health implications for the individual and for society[23].

[22] For the main ethical issues raised by the diffusion of ICT, cfr. National Committee for Bioethics, *Information and Communication Technologies and Big Data: Bioethical Issues*, 25 November 2016, http://bioetica.governo.it/media/3207/p124_2016_information-techonolgies-and-big-data_en.pdf (in particular, pp. 13-16).

[23] In this respect, the decrease in immunization coverage is a good example because the tendency to defer or refuse vaccinations has consequences at an individual and collective level, for example invalidating the protection of vulnerable people, including those who cannot be

6. *The risk of the boomerang effect*

At this point, the following question could be raised: what should medicine do in order to solve the "Cassandra problem", that is, to face «unwarranted suspicion and misjudged refusal to trust, even where there is adequate – if inevitably imperfect – evidence of trustworthiness» with regard to medical knowledge[24]? Surely, the adequate promotion of competence criteria is required in order to acknowledge and accept its peculiar traits. For example, differently from the power of prophecy possessed by the mythological figure of Cassandra, medical competence is not a gift, but rather the result of an endless studying and training process. This is why not everyone can be considered an expert in medicine, although medical knowledge might be obtained, at least in principle, by all. In other words, competence criteria show an inherent selective character and not all different opinions on scientific matters are equally valuable. Medical competence assigns a stronger force to experts' opinions, which are however called to be continuously discussed within the medical scientific community. Once again, it is precisely the revisability of medical knowledge and, in parallel, intrinsic uncertainty in medicine that guarantee the authoritativeness of expert medical opinions and competences.

The question that has been proposed at the beginning of this paragraph could thus be reformulated in the following terms: what should medicine do in order to show that medical knowledge is not absolute but, despite the presence of intrinsic uncertainty, is nonetheless characterized by a certain degree of certainty? When illustrating both discoveries in the medical field and medical advances that are expected for the future, a careful selection and use of words by the medical scientific community would be a good starting point. Terms should not be ambiguous, for example, suggesting the highest degree of certainty and thus obscuring the irreducible uncertainty of medicine. Actually, some words currently used in the era of evidence-based medicine (EBM) and precision medicine (PM) would seem to assign a fully scientific character to medicine. For example, at first glance, "evi-

vaccinated for health reasons. A greater personal and social responsibility should thus be assumed, and falsehoods regarding vaccines (such as the scientifically unfounded idea that vaccination triggers autism) should be dispelled.

[24] O. O'Neill, *Autonomy and Trust in Bioethics*, Cambridge University Press, Cambridge 2005[3], pp. 141-142 (quotation is at p. 141). In Greek mythology, Cassandra was a daughter of Priam, the King of Troy, who received the gift of prophecy from Apollo. Nevertheless, when Cassandra refused Apollo's love, he condemned her to never be believed to the extent that, despite her trustworthiness, her prophecy regarding the fall and destruction of Troy went unheeded.

dence" could be considered as an absolute category, entirely free of context[25]. In this way, in reference to EBM's main instruments, i.e. clinical practice guidelines and research protocols, EBM might be erroneously expected to rely on indisputable facts. Besides, "evidence" derives from the Latin *evidentia* which means vividness or clearness; when literally translated into other languages this word can mainly lead one to consider scientific evidence as an irrefutable form of knowledge. For instance, the literal Italian translation of "evidence", which is "evidenza", alludes to something that cannot be questioned or denied precisely because of its clearness.

The previous remarks do not intend to question the contribution of EBM[26]. Information obtained by randomized controlled trials (RCTs), and systematic reviews of RCTs as well, regarding the efficacy and safety of healthcare interventions provides clinical practice guidelines, such as step-by-step instructions and estimates of the treatment outcomes, usually helping to reach medical decisions. Moreover, adequate communication of evidence results to patients can lead to a more shared decision-making process because, for example, statistical data may enhance patients' participation in the context of discussing risk. Nevertheless, the evidence provided by EBM concerning the effectiveness of interventions might obscure the intrinsic uncertainty of medicine because, at first glance, the word "evidence" alludes to something that is indisputably evident, and thus totally certain.

Similar considerations can be made for the word "precision" because, especially in the colloquial sense, it «implies a high degree of certainty of an outcome, as in "precision-guided missile" or "at what precise time will you arrive?»[27]. The terminological ambiguity seems to be further emphasized by the emerging concept of systems medicine, which is often promoted as "P4 medicine" (predictive, preventive, personalized and participatory) and conveys advances in genetic research towards PM[28]. Indeed, focused on the

[25] This risk has been addressed, for example, by the Canadian Health Services Research Foundation within a document that summarizes the results of a workshop held in 2005 and focused on scientific evidence. See Canadian Health Services Research Foundation, *Weighing Up the Evidence. Making Evidence-Informed Guidance Accurate, Achievable, and Acceptable*, January 2006, https://www.cfhi-fcass.ca/migrated/pdf/weighing_up_the_evidence_e.pdf.

[26] For the main advantages of EBM methods, see W. Rogers, K. Hutchison, *Evidence-Based Medicine in Theory and Practice: Epistemological and Normative Issues*, in Schramme, Edwards (eds.), *Handbook of the Philosophy of Medicine*, cit., pp. 851-872 (in particular, pp. 852-857).

[27] D.J. Hunter, *Uncertainty in the Era of Precision Medicine*, in «The New England Journal of Medicine», 375 (2016), n. 8, p. 711.

[28] For an overview of advances in the field of PM as well as of their implications, see H.-P. Deigner, M. Kohl (eds), *Precision Medicine: Tools and Quantitative Approaches*, Academic Press,

intersection of three factors, i.e. individual variations in genes, environmental interactions and influence of lifestyle, the four Ps are associated with promises of a forthcoming revolution in medicine[29]. Systems medicine pledges to provide predictive assessments, that is, individual health risk information relating to potential future genetic diseases, and thus to facilitate prevention, personalize medicine and motivate people to change their health-related behavior, reducing the risk of the disease's onset[30].

By illustrating medicine as a scientific enterprise able to achieve these goals in the short or long term, a boomerang effect could occur. The four Ps might present medicine itself as a science and a practice characterized by full certainty, fostering in this way claims of infallibility by the individual and society on the one hand, and suspicions as well as incredulity on the other hand. In other words, such a boomerang effect might negatively affect the trust-medicine dyad because medicine itself could lead to a refusal of its epistemological status or paradoxically contribute to a decrease in its authoritativeness, respectively promoting displaced trust and displaced distrust in medicine.

7. *The certainty of increasing uncertainty*

Beyond the words that are chosen to illustrate medical advances, benefits of EBM and developments in genetic testing technology could constitute

Amsterdam 2018; M. Verma, D. Barh (eds), *Progress and Challenges in Precision Medicine*, Academic Press, London 2017.

[29] Cfr. L. Hood, R. Balling, C. Auffray, *Revolutionizing Medicine in the 21st Century through Systems Medicine*, in «Biotechnology Journal», 7 (2012), n. 8, pp. 992-1001. See also M. Flores, G. Glusman, K. Brogaard, N.D. Price, L. Hood, *P4 Medicine: How Systems Medicine Will Transform the Healthcare Sector and Society*, in «Personalized Medicine», 10 (2013), n. 6, pp. 565-576.

[30] In this respect, the growing diffusion of direct-to-consumer (DTC) genetic susceptibility tests should not be underestimated. This kind of testing can be purchased at increasingly reduced prices and gives everyone the possibility to obtain health risk information without the guidance or supervision of healthcare providers or genetic counselors. It is not possible to discuss here the main problematic issues raised by these tests. The following remark is only mentioned as related to the focus of this paper. Although research is ongoing, genetic susceptibility testing has limited predictive power because it carries a degree of uncertainty as to whether a disease will develop, when it will develop and how severe it will be. The use of DTC genetic susceptibility tests usually increases the problem of managing the impact of this uncertainty. Indeed, without specialized knowledge of genetics and the involvement of healthcare professionals, misinterpretation of test results may occur and entail unjustified negative feelings, such as anxiety and depression. For an in-depth analysis of the main ethical issues of DTC genetic susceptibility tests as well as of their implications for healthcare systems, see F. Marin, *Putting Health in the Marketplace. Ethical Issues about Providing Online Health Risk Information*, «Medicina e Morale», 66 (2017), n. 1, pp. 31-43.

themselves a proof that a decrease, or even an elimination, of uncertainty is really possible. For example, considering the ability of well-designed RCTs to inform medical practice and to guide the decision-making process as well as the current possibility to scan and compare entire genomes, a reduction of medical uncertainty may intuitively be expected in EBM and PM.

Actually, advances in these fields, rather than diminishing medical uncertainty, are contributing to its increase and even generating new varieties of uncertainties. Indeed, as regards EBM, it has been shown that, given the increased reliance on information technologies and epidemiology, research protocols and evidence-based guidelines generate new kinds of uncertainties. For example, through in-depth interviews with pediatric residents from two medical programs about their experiences with EBM, Stefan Timmermans and Alison Angell have addressed a new form of uncertainty, named "research-based uncertainty"[31]. Some of the interviewees felt uncomfortable in conducting literature searches or evaluating protocols and guidelines, while others felt unsure in distinguishing a good sample from a bad one as well as in differentiating statistical significance from confidence intervals.

A greater set of biomedical and epidemiological variables is offered by PM as well[32]. Indeed, quoting Lily Hoffman-Andrews, «the wonderful promise of expanding testing has, in practice, run into the frustrating reality of a greater burden of uncertain results»[33]. For example, when reviewing thousands of variants obtained by exome sequencing, clinicians and laboratory personnel deal with many variants of uncertain significance (VUSs) and have to decide whether to omit or include them[34]. When encountering VUSs, troublesome questions concern the care relationship as well because genetics professionals and clinicians are asked whether and how to disclose these variants to the patient and how to manage their impact[35].

[31] S. Timmermans, A. Angell, *Evidence-Based Medicine, Clinical Uncertainty, and Learning to Doctor*, in «Journal of Health and Social Behaviour», 42 (2001), n. 4, pp. 342–359 (in particular, pp. 348-349). A reworked version of this article is chapter 5 of a book that Stefan Timmermans has written with Marc Berg. See S. Timmermans, M. Berg, *The Gold Standard: The Challenge of Evidence-Based Medicine and Standardization in Health Care*, Temple University Press, Philadelphia 2003, pp. 142-165.

[32] D.J. Hunter, *art. cit.*, pp. 711-713.

[33] L. Hoffman-Andrews, *The Known Unknown: The Challenges of Genetic Variants of Uncertain Significance in Clinical Practice*, in «Journal of Law and the Biosciences», 2017, pp. 648-657 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5965500/pdf/lsx038.pdf), quotation is at p. 649.

[34] S. Timmermans, C. Tietbohl, E. Skaperdas, *Narrating Uncertainty: Variants of Uncertain Significance (VUS) in Clinical Exome Sequencing*, in «BioSocieties», 12 (2017), n. 3, pp. 439-458.

[35] See K. Barlow-Stewart, *The Certainty of Uncertainty in Genomic Medicine: Managing the Challenge*, in «Journal of Healthcare Communication», 3 (2018), n. 3, p. 37.

As a consequence, although the expansion of medical knowledge is expected to reduce informational uncertainty, the certainty of increasing medical uncertainty is confirmed properly by medical progress and technological innovations. This is why, nowadays, dealing with the medicine-uncertainty dyad requires greater tolerance of uncertainty and further strategies to face new kinds of uncertainties derived by medical advances.

8. *Conclusions: from a twofold dyad to a triple pattern*

This paper has not merely addressed the crucial role of trust in medical practice and the ubiquitous presence of uncertainty in medicine, as tends to happen in scientific literature; rather, it has gone further by showing that several problematic issues arise when the trust-medicine dyad is recognized without the acknowledgment of the medicine-uncertainty dyad, or vice versa. Indeed, displaced trust and displaced distrust in medicine occur when respectively considering medical knowledge as an absolutely certain knowledge and refusing uncertainty as an inherent feature of medicine. In this way, the main thesis of the paper has been justified, that trust in medicine is well-placed when the trust-medicine-uncertainty interdependency is fully recognized and adequately valued.

The triple pattern proposed in these pages is particularly advantageous for the following reasons. Firstly, it adds value to the epistemological status of medicine because intrinsic uncertainty is proof of the revisable character of medical enterprise, and such revisability guarantees a well-placed trust in medicine. Secondly, the promotion of the trust-medicine-uncertainty interdependency involves healthcare professionals, patients, mass media, and society in general. Indeed, everyone is called to recognize the authoritativeness of medical knowledge and competence criteria as well as to paradoxically appreciate episodes of dissent within the scientific community as proof of the revisability of medicine. Thirdly, the triple pattern proposes a sort of balancing between certainty and uncertainty. By admitting irreducible uncertainty in medicine, the authoritativeness of medical knowledge as well as the contribution of medical-technological progress are not questioned and at the same time an overestimation of uncertainty does not occur. Instead, the interdependency mentioned above denies the ascription of a fully scientific character to medicine on one side, and an antiscientific view of medical knowledge on the other side.

As far as the final point is concerned, is must be noted that balancing

certainty and uncertainty avoids a twofold risk. On the one hand, absolutizing certainty leads healthcare professionals to take less responsibility for their actions, negatively affecting the patient's health. On the other hand, an overemphasis on uncertainty leads to over-responsibility, questioning any plan of care on behalf of the patient. In this way, medicine as a science and a practice is dangerously considered as a world which is neither white nor black. Actually, the certainty of uncertainty makes medicine a gray-scale space[36] toward which, as it has been argued in this paper, trust can be well-placed by precisely dealing with medical uncertainty. In other words, adequately facing the challenges posed by this grey-scale scenario means recognizing and promoting the trust-medicine-uncertainty interdependence.

Abstract

The paper does not merely address the crucial role of trust in medical practice and the ubiquitous presence of uncertainty in medicine, as tends to happen in scientific literature; rather, it goes further by showing that problematic issues arise when the trust-medicine dyad is recognized without the acknowledgment of the medicine-uncertainty dyad, or vice versa. Firstly, it is argued that the trust-medicine-uncertainty interdependency is necessary because there is a kind of irreducible uncertainty due to the epistemological status of medicine, whose presence guarantees well-placed trust in medicine. *In this respect, examples of* misplaced trust in medicine *due to considering medicine as an absolutely certain scientific knowledge and* misplaced distrust in medicine *as a result of an antiscientific view of medical knowledge are discussed. Secondly, the need for a triple pattern is proved to be urgent because medical advances, rather than diminishing medical uncertainty, are contributing to its increase and even generating new kinds of uncertainties.*

Keywords: trust; medical practice; medical uncertainty; evidence-based medicine; precision medicine.

Francesca Marin
Università degli Studi di Padova
francesca.marin@unipd.it

[36] A.L. Simpkin, R.M. Schwartzstein, *art. cit.*, p. 1714.

T

L'esemplarità, inattuale proposta di senso esposta alla prova della fiducia. Un'analisi a partire da Bergson e Scheler

Maria Teresa Russo

1. *Esemplarità e fiducia, categorie attuali e inattuali*

L'odierna crisi del legame sociale, presentata e analizzata da svariate indagini, rende oggi più attuale la questione della relazione di fiducia, reclamata in diversi ambiti, da quello finanziario a quello medico[1]. La necessità di incrementare il cosiddetto "capitale sociale", strutturato attorno alla cooperazione e alla solidarietà, nonché la centralità che temi come il dono e la gratuità stanno assumendo anche in economia, oltre all'importanza assegnata nella ricerca pedagogica a concetti come patto di corresponsabilità e alleanza educativa, portano in primo piano il bisogno di recuperare aspetti che il lato oscuro dell'individualismo ha messo progressivamente in ombra. Si parla persino di un'"etica del vicinato"[2], che superi l'indifferenza reciproca o la sensazione di minaccia che lo spaesamento delle grandi città contribuisce ad alimentare[3].

La fiducia appare quindi come il correttivo più richiesto in un mondo

[1] Cfr. ad esempio O. O' Neill, *Autonomy and Trust in Bioethics*, Cambridge University Press, Cambridge 2002.

[2] Cfr. H. L'Heuillet, *Du voisinage. Réflexions sur la coexistence humaine*, Albin Michel, Paris 2016.

[3] Sono numerose le indagini recenti di sociologia urbana e di psicologia sociale che analizzano le dimensioni del sentimento di insicurezza, nonché i fattori psicologici che influenzano la percezione del rischio nel contesto delle città moderne. Si veda: B. Zani (a cura di), *Sentirsi in/sicuri in città*, il Mulino, Bologna 2003; R. Sennett, *The Rituals, Pleasures and Politics of Cooperation,* Yale University Press, New Haven, USA 2012. Trad. di A. Bottini, *Insieme. Rituali, piaceri, politiche della collaborazione*, Feltrinelli, Milano 2012; e anche R. Sennett, *Building and Dwelling: Ethics for the City*, Farrar, Straus & Giroux, New York 2018. Trad. di C. Spinoglio, *Costruire e abitare. Etica per la città*, Feltrinelli, Milano 2018.

 DOI: 10.4454/teoria.v39i1.59

che ha reso tutti più vicini, grazie alla rapidità dei mezzi di comunicazione e di informazione, ma contemporaneamente tutti più lontani, a causa della precarietà e dell'isolamento che questi stessi mezzi possono generare. Le stesse risorse tecnologiche finalizzate alla sicurezza, se divengono strumenti di controllo sociale, finiscono per aumentare il sospetto e la diffidenza reciproca[4]. Sorge allora più prepotente, oltre al bisogno di riconquistare gli spazi di autonomia perduti, la nostalgia per modalità di relazione improntate alla prossimità fiduciosa riducendo la distanza sociale[5].

D'altra parte, è giustificato parlare anche di una inattualità della fiducia. Come osserva Linda Zagzebski[6], questa radica nella crisi oggi sperimentata del concetto stesso di autorità epistemica, ossia di una persona degna di fiducia in quanto esperta in un determinato ambito. La crisi dell'"esperto" affidabile si manifesta sia in quanto i settori di competenza sono spesso molto ristretti e in scarso rapporto gli uni con gli altri, sia in quanto sui temi valoriali, politici o religiosi non si accetta più la guida di un'autorità. Il motivo teorico più importante di questa crisi sta, secondo Zagzebski, nel conflitto tra l'autorità epistemica e due valori tipicamente moderni: l'egualitarismo e l'autonomia. Quando quest'ultima si declina secondo il paradigma individualista dell'autosufficienza e dell'autorealizzazione[7], per il quale la relazione con l'altro risulta superflua o persino controproducente, lo spazio della fiducia si erode necessariamente[8]. Appare

[4] Cfr. A. Greenfield, *Radical Technologies: The Design of Everyday Life*, Verso Books, London-New York 2018. Trad. di M. Nicoli, A. Manna, M. Ferrara, C. Veltri, *Tecnologie radicali. Il progetto della vita quotidiana*, Einaudi, Torino 2017.

[5] Si veda, ad esempio, il fenomeno delle "Social Street", costituite da gruppi di vicini che si accordano per costituire punti di incontro, conoscersi e collaborare. C. Pasqualini (a cura di), *Vicini e connessi. Rapporto sulle Social Street a Milano*, Fondazione Feltrinelli, Milano 2018.

[6] L.T. Zagzebski, *Epistemic authority: a theory of trust, authority, and autonomy in belief*, Oxford University Press, Oxford-New York 2012, pp. 5-6.

[7] Cfr. Ch. Taylor, *The Malaise of Modernity*, Anansi, Concord (Ont.) 1991.Trad. di G. Ferrara degli Uberti, *Il disagio della modernità*, Laterza, Roma-Bari 1999.

[8] «L'individualismo nega la partecipazione mediante l'isolamento della persona intesa solo come individuo e concentrata su se stessa e sul suo proprio bene, che viene pure concepito come isolato dal bene degli altri e anche dal bene comune. Il bene dell'individuo ha, in questa concezione, carattere addirittura contrapposto ad ogni altro individuo e al suo bene, e, comunque, carattere di "autoconservazione" e difensivo. L'agire insieme con gli altri, come l'esistere insieme con gli altri, è, secondo l'individualismo, una necessità cui l'individuo deve piegarsi, ma a questa necessità non corrisponde alcuna qualità positiva dell'individuo, e l'agire e l'esistere insieme con gli altri non servono e non sviluppano nessuna di tali qualità. Gli "altri" sono per l'individuo solo fonte di limitazione e perfino polo di molteplici contrasti. La comunità quando sorge ha come scopo quello di assicurare il bene dell'individuo in mezzo agli "altri". Ecco, in breve, delineata la posizione individualistica». K. Wojtyla, *The acting person*, D. Reidel Publishing

all'orizzonte il paradigma dell'immunizzazione, che pretende di proteggersi dalla "ferita dell'altro"[9], da quella vulnerabilità che si manifesta proprio nel bisogno di fidarsi e di affidarsi.

Anche la centralità assegnata alla fiducia nell'ambito di una certa "etica femminista", come quella teorizzata da Annette Baier[10], non è esente dall'ambiguità di una interpretazione emotivista. Se è valido l'invito a non sottovalutare il peso nella condotta morale dei cosiddetti "pregiudizi morali", ossia di quella dimensione sentimentale e desiderativa a cui appartiene anche la fiducia, resta tuttavia da discutere se tale valorizzazione debba avvenire a discapito dei giudizi e delle motivazioni razionali.

Eppure, laddove è in gioco l'esercizio di un'autorità o comunque una relazione asimmetrica, sia che si tratti del rapporto tra maestro e allievo, tra medico e paziente o tra politico e cittadino, solo la fiducia è una garanzia contro la deriva della prevaricazione o della violenza[11].

Un discorso analogo vale anche per il concetto di esemplarità, categoria che allo stesso modo può essere considerata attuale e inattuale. Da un lato cresce il desiderio di emulazione e di modelli credibili, che possano fornire uno sprone soprattutto alle giovani generazioni, dall'altro la cosiddetta "società adiaforica" non sembra propensa a fornire una segnaletica per la strada valoriale da percorrere[12]. Per di più, a causa dell'estetizzazione della nostra cultura, è sempre più spesso la società dello spettacolo e dell'apparire a suggerire una esemplarità priva di contenuti etici ma, allo stesso

Company, Dordrecht 1979. Trad. di S. Morawski, R. Panzone, R. Liotta, *Persona e Atto*, Libreria Editrice Vaticana, Città del Vaticano 1982, pp. 310-311.

[9] Cfr. L. Bruni, *La ferita dell'altro. Economia e relazioni umane*, Il Margine, Trento 2007.

[10] Annette Baier definisce la fiducia come «the accepted vulnerability to another's possible but not expected ill will (or lack of good will) toward one». A. Baier, *Trust and Anti-Trust*, in «Ethics», vol. 96, n. 2 (Jan., 1986), p. 235; Ora in Id., *Moral Prejudices: Essays on Ethics*, Harvard University Press, Cambridge 1995, p. 152.

[11] Con parole di Ricoeur: «proprio il legame di natura fiduciaria fa la differenza ultima fra autorità e violenza nel cuore stesso del rapporto gerarchico di dominazione». P. Ricoeur, *Le paradoxe de l'autorité,* in *Le Juste 2,* Esprit, Paris 2001, pp. 107-123. Trad. di D. Iannotta, *Il paradosso dell'autorità*, in P. Ricoeur, *Il Giusto 2*, Effatà, Cantalupa (To) 2007, p. 131. Si veda anche: H. Arendt, *What Is Authority?* (1961), in *Between Past And Future: Eight Exercises in Political Thought*, Penguin Books, New York 2006, pp. 91-141. Trad. di T. Gargiulo, *Cosa è l'autorità*, in Id., *Tra passato e futuro*, Garzanti, Milano 1999., pp. 130-192. Ancora la Arendt ritiene che tanto la politica, come la legge e i contratti esistano per garantire un minimo di fiducia: «creano un quadro di prevedibilità all'interno dell'imprevedibile». H. Arendt, *Denktagebuch: 1950-1973*, Piper Verlag, München 2002. Trad. di C. Marazia, *Quaderni e diari 1950-1973*, Neri Pozza, Vicenza 2007, pp. 116 ss.

[12] Cfr. Z. Bauman, *Postmodern Ethics*, Blackwell, Oxford 1993. Trad. di G. Bettini, *Le sfide dell'etica*, Feltrinelli, Milano 1996.

tempo, fortemente seduttiva[13], sebbene i modelli proposti abbiano vita ogni volta più breve e siano smitizzati con la stessa rapidità con cui erano stati creati[14]. L'esemplarità conserva dunque il suo potere di attrazione, ma è sempre più difficile distinguere tra esemplarità e controesemplarità.

Su quali presupposti allora si genera la fiducia in un modello e fino a che punto questi è affidabile?

È noto come per Kant non si dia l'imitazione di modelli[15], che egli ritiene dannosi sia moralmente, perché comportano una dipendenza dal sensibile, sia anche teoreticamente, in quanto o favoriscono generalizzazioni arbitrarie o, al contrario, inducono a concentrarsi indebitamente sul particolare. Il valore è dunque rappresentato dalla norma o legge che è un a priori di cui la persona è espressione, Anche nelle *Lezioni di etica*, tenute tra il 1775 e il 1781, a proposito della conoscenza morale e religiosa, egli ribadisce che «si manifesta apoditticamente a priori per mezzo della ragione, cioè noi riconosciamo a priori la necessità di comportarci in un certo modo e non altrimenti; perciò, in materia di morale e di religione, non è necessario alcun esempio». Per questo motivo, a suo avviso l'esempio di una vita buona o santa potrà anche essere utilizzato per incoraggiare, ma non certo indicato come modello[16].

Di diversa opinione è Croce, il quale contesta chi nega l'efficacia

[13] Cfr. R. Bubner, *Ästhetische Erfahrung*, Suhrkamp, Frankfurt/Main 1989. Trad. di M. Ferrando, *Esperienza estetica*, Rosenberg & Sellier, Torino 1992, pp. 109-116.

[14] Si veda il 15° Rapporto Censis sulla Comunicazione, *I media digitali e la fine dello star system*, Roma, 11 ottobre 2018, dove si evidenzia il declino del divismo, ossia di quel pantheon di idoli ed "eroi", figure simboliche che incarnavano un modello di vita migliore e desiderabile: «oggi la moltitudine dei soggetti, novelli Prometeo dell'era digitale, ha trascinato quel pantheon giù dall'Olimpo nel disincanto del mondo. Uno vale un divo: siamo tutti divi. O nessuno, in realtà, lo è più». Disponibile online: www.censis.it

[15] «In sede morale non c'è posto per l'imitazione, e gli esempi non servono che da incoraggiamento, cioè a togliere ogni dubbio sulla ottemperabilità di ciò che la legge comanda, a rendere intuibile ciò che la regola pratica esprime in modo più generale, ma non è ammissibile che sia posto in disparte il loro vero originale, che si trova nella ragione, e che ci si regoli su esempi». I. Kant, *Grundlegung zur Metaphysik der Sitten,* in *Gesammelten Schriften. Erste Abtheilung: Werke*. Band IV, Herausgegeben von der Akademie der Wissenschaften Königlich Preussischen, Georg Reimer, Berlin 1911, pp. 385-464. Trad. di P. Chiodi, *Fondazione della metafisica dei costumi. Passaggio dalla filosofia morale popolare alla metafisica dei costumi*, Utet, Torino 2013, p. 45.

[16] «Quando perciò in religione ci viene indicata come modello la santità di alcune persone, siano esse sante quanto si vuole, io non debbo imitarle ma valutarne il contegno alla luce dei principi universali della condotta. […] Le valuto raffrontandole alla santità della legge: solo quando esse si accordano con questa, io riconosco che si tratta di esempi di santità». I. Kant, *Eine Vorlesung Kants über Ethik*, in *Auftrage der Kantgesellschaft,* Pan Verlag Rolf Heise, Berlin 1924, pp. VII-335. Trad. di A. Guerra, in *Lezioni di etica*, Laterza, Roma-Bari, 1971, pp. 126-128.

dell'esempio adducendo come motivo che la forza morale non può essere risvegliata da fattori esterni se non la si possiede. Invece per il filosofo l'esempio di santi ed eroi influisce in modo decisivo sulla condotta, quando la sfiducia o il dubbio suggeriscono di cedere o di arrendersi[17]. Nella sua prospettiva, la forza morale è quasi ipostatizzata, in quanto non è una proprietà dell'individuo, ma lo trascende: questo è il motivo per cui non può venir meno e l'esempio ne è una determinazione particolare. All'opposto, è innegabile «l'efficacia depressiva, sottilmente corruttrice, lentamente devastatrice»[18] del cattivo esempio di azioni che finiscono per avere vita autonoma, staccata da chi le ha compiute, il cui effetto è quello di «aggravare la sfiducia dell'uomo operante e lottante»[19].

Da qui il dubbio che non possa esistere una esemplarità autentica in un contesto relativistico. Rorty, con la sua "filosofia edificante" che considera la persuasività di un'affermazione o di un esempio sempre solo relativa a un orizzonte, senza alcuna pretesa di validità universale, giacché dire il contrario sarebbe affermare un uso autoritativo della razionalità[20], smantella l'idea che l'esemplarità filosofica sia vincolata a un concetto oggettivo di verità. In questa prospettiva, la fiducia diventa un surrogato della veridicità: mi fido di te, non perché dici o pratichi la verità, ma perché trovo in te una certa coerenza o congruenza, indipendentemente dal riferimento a un principio oggettivo che ti oltrepassa, a un orizzonte di senso che superi il modello[21].

In questa prospettiva, il discorso esemplare si legittimerebbe per l'effettualità e la performatività del suo stesso sviluppo. In altri termini, il modello non sarebbe tale in quanto portatore di valore o di verità, ma in quanto innesca un movimento nel discepolo. Insomma, sarebbe la fiducia a fare il modello e non il contrario: non si crede né alla persona, né a ciò

[17] Cfr. B. Croce, *L'efficacia dell'esempio*, in *Etica e politica*, Laterza, Roma-Bari 1956, pp. 149-154.

[18] *Ivi*, p. 153.

[19] *Ibidem*.

[20] «A mio parere, la nozione di "validità universale" e quella di "corrispondenza a una realtà indipendente" sono solidali: se cade l'una, cade anche l'altra. Non perdiamo nulla se smettiamo di utilizzare entrambe. Allora la distinzione tra ciò che è giustificato per noi e ciò che è vero può essere sostituita dalla distinzione tra essere in grado di giustificare le nostre credenze a certi ascoltatori e non a certi altri». R. Rorty, *Le asserzioni sono pretese di validità universale?*, in G. Vattimo (a cura di), *Filosofia '94*, Laterza, Roma-Bari 1995, p. 59.

[21] Analoga è anche la posizione di A. Ferrara, che argomenta l'efficacia dell'esempio in una prospettiva non fondazionalista dell'etica, dove «possedere esemplarità vuol dire essere legge a se stessi», mostrando un'autocongruenza eccezionale che produce arricchimento ed espansione, al pari di un'opera d'arte. A. Ferrara, *La forza dell'esempio. Il paradigma del giudizio*, Feltrinelli, Milano 2008, pp. 107-108.

che dice, ma alla relazione che si crea, una sorta di adesione per "analogia vitale"[22]. Non è più pertanto in gioco una funzione maieutica e pedagogica del modello, che presuppone una componente oggettiva dell'esemplarità, in quanto – si afferma – «la sua oggettività non consiste ora più nell'apprendimento di un senso dato, ma più precisamente in un senso che sfugge a ogni soggettività, persino a quella del nuovo "maestro" di verità, e che in qualche misura si mantiene oscillante tra l'arcano e l'inconscio»[23].

In tale visione, dove ad acquistare forza è la fedeltà all'alterità e non la fedeltà alla verità incarnata nell'altro, il vecchio adagio "amicus Plato sed magis amica veritas", viene rovesciato nel suo contrario. Afferma Vattimo:

> Il pensiero che non si concepisce più come riconoscimento e accettazione di un fondamento oggettivo perentorio svilupperà un nuovo senso della responsabilità, come disponibilità e capacità, alla lettera, di rispondere agli altri da cui, in quanto non fondato sull'eterna struttura dell'essere, si sa "proveniente". *Amica veritas, sed magis amicus Plato*, forse[24].

Invece l'esemplarità e, il suo correlato, la fiducia, non possono prescindere da quella che Ricoeur denomina "una teleologia soggiacente"[25], ossia dal riferimento a una universalità di valori, che consenta di distinguere tra vita autentica e inautentica. Non è un caso che il tedesco *Vorbild*, guida, modello, risulti composto da *Bild*, immagine o forma dell'uomo autentico, a cui a sua volta la *Bildung*, il processo formativo, deve mirare. Il modello è in qualche modo maestro di vita, l'incarnazione di una qualche verità. La nozione di "maestría" non può mai essere tale solo grazie alle virtù intellettuali o ai contenuti di quanto si insegna, ma per ciò che si mostra con la propria condotta.

Chiamiamo, infatti, coerenza proprio quella continuità e quell'adeguatezza tra i giudizi formulati da qualcuno e le sue azioni, che non soltanto

[22] «Quando si abbandonano l'errore e la falsità, è perché si revoca la propria fiducia a certi maestri e la si concede ad altri. Può succedere che uno trovi intorno a sé solo persone che meritano sfiducia, o che viceversa emerga solo in lui stesso la ragionevole pretesa di essere creduto dagli altri… Per questa ragione la virtù della credenza razionale non è la fede in qualcosa che non vediamo ma la fiducia negli altri, in *qualcun* altro». Lluis Alvárez, *La filosofia come fiducia*, in G. Vattimo (a cura di), *Filosofia '94*, cit., pp. 76-77.

[23] *Ivi*, p. 92.

[24] G. Vattimo, *Oltre l'interpretazione. Il significato dell'ermeneutica per la filosofia*, Laterza, Roma-Bari 1994, p. 52.

[25] Cfr. P. Ricoeur, *Jugement esthétique et jugement politique selon Hannah Arendt*, in *Le Juste*, 1, Esprit, Paris 1995, pp. 143-161. Trad. di D. Iannotta, *Giudizio estetico e giudizio politico in Hannah Arendt*, in P. Ricoeur, *Il Giusto*, SEI, Torino 1995, p. 134.

ne rendono la condotta comprensibile, favorendo le relazioni interpersonali, ma conferiscono credibilità e hanno una forza esemplare. È significativo l'episodio riportato da Gadamer, che mostra la sorpresa di chi riscontra una mancanza di coerenza in un intellettuale considerato un maestro:

> Max Scheler, il fondatore dell'etica materiale del valore, dette un giorno la seguente risposta ad un alunno che gli chiedeva conto del perché egli descrivesse così chiaramente l'ordine dei valori e la loro forza normativa e poi vi si adeguasse ben poco nella sua vita: "Va forse il cartello nella direzione che esso stesso indica?"[26].

Inoltre, nell'esemplarità autentica l'investimento fiduciario non genera dipendenza passiva, ma fioritura reale. In caso contrario, sarebbe una forma larvata di potere. Il modello è "edificante", nel senso che induce in chi vi si ispira alla costruzione di elementi che senza il suo esempio non sarebbero possibili.

2. *La fiducia nelle società chiuse e la causalità esemplare in Henri Bergson*

Com'è noto, nell'opera forse più discussa di H. Bergson, *Les deux sources de la morale et de la religion*, la questione della «causalità esemplare» viene posta nei termini di un interrogativo: «Perché i grandi uomini di bene hanno trascinato (*ont entraîné*) dietro di loro le folle?»[27]. È interessante soffermarsi sul termine «entraîner»[28], che ricorre ben 23 volte ne *Les Deux Sources*, tanto da qualificare tali modelli come i grandi «entraîneurs de l'humanité»[29]. Ma mentre in diversi saggi dell'epoca, il verbo «entraîner»

[26] H.-G. Gadamer, *Über die Möglichkeit einer philosophischen Ethik* (1963), *Neuere Philosophie II: Probleme, Gestalten, in Gesammelte Werke* 4, J.C.B. Mohr (Paul Siebeck), Tübingen 1987, pp. 175-188. Trad. di U. Margiotta, *Sulla possibilità di un'etica filosofica, in Ermeneutica e metodica universale*, Marietti, Torino 1973, p. 147.

[27] H. Bergson, *Les deux sources de la morale et de la religion*, Puf, Parigi 2008, p. 30. Ci si riferirà sempre a questa edizione critica, curata da F. Worms.

[28] Il *Dictionnaire Larousse* segnala diversi significati che mostrano la ricchezza semantica del verbo: «Faire connaître à quelqu'un le même état, la même évolution que soi-même; conduire quelqu'un avec soi quelque part, l'amener de force; amener quelqu'un, par une pression morale, par la séduction ou l'exemple, à agir, à s'engager dans une voie qu'il n'a pas délibérément choisie; pousser: amener tel comportement de la part de quelqu'un, en être la cause, avoir tel résultat, telle conséquence; impliquer, engager; exercer un effet stimulant sur quelqu'un, le pousser irrésistiblement».

[29] H. Bergson, *Les deux sources de la morale et de la religion*, cit., p. 55.

designa il più delle volte un processo in cui non tanto l'individuo, ma la folla è passivamente determinata ad agire per una sorta di suggestione, quindi anche in modo incosciente[30], in Bergson il termine indica piuttosto l'apertura fiduciosa e consapevole da parte del singolo individuo al richiamo di una "grande personalità morale"[31]. Sono queste figure a imprimere una svolta nelle relazioni sociali, dando luogo alla trasformazione da una società in cui la sicurezza è prodotta dall'osservanza di obblighi e di riti a un'altra, in cui è l'amore il fattore più potente di coesione.

Per il filosofo, l'essere umano, nel momento in cui diviene consapevole della sua vulnerabilità, è essenzialmente bisognoso di speranza e di fiducia. "A difetto di potenza, abbiamo bisogno di fiducia"[32]: costruirci una rassicurazione nei confronti della funzione di dissoluzione, ossia di scomposizione analitica del reale, esercitata dall'intelligenza, che mostra inesorabilmente la nostra mortalità, generando angoscia e paura[33]. È da questa minaccia che la natura si difende, opponendovi la credenza, che «significa dunque essenzialmente fiducia»[34]: una rassicurazione contro il timore e lo scoraggiamento indotti dalla conoscenza intellettuale della propria fragilità e dell'impossibilità di controllare il futuro. Questo timore è anche la causa della rottura della coesione sociale, in quanto origina nel singolo la necessità egoistica di preservare se stesso, nutrendo diffidenza nei confronti dell'altro.

Per fare da contrappeso a tale reazione, sorge la cosiddetta "funzione fabulatrice" – un residuo d'istinto vitale – che ispira ai membri della società una fiducia fondamentale nella vita e nella natura, dando luogo a una forma statica di religione, strutturata attorno a proibizioni e obblighi e a una società chiusa, preoccupata di difendere gli interessi del gruppo[35]. Nella società chiusa, la disciplina assicura coesione interna e difesa dalle minacce esterne: supportata dalla funzione fabulatrice della religione, mette in atto un meccanismo che all'interno consiste nel "correggere" gli effetti della

[30] Si vedano, ad esempio le indagini di Gustave Le Bon, *Psychologie des foules* (1895) o di Gabriel Tarde, *L'opinion et la foule* (1901).

[31] H. Bergson, *Les deux sources de la morale et de la religion*, cit., p. 30.

[32] *Ivi*, p. 172.

[33] «L'uomo non può esercitare la sua facoltà di pensare senza rappresentarsi un futuro incerto, che risveglia il suo timore e la sua speranza». *Ivi*, p. 216.

[34] *Ivi*, p. 159.

[35] «Una società chiusa non può vivere, resistere all'azione dissolutrice dell'intelligenza, conservare e comunicare a ciascuno dei suoi membri la fiducia indispensabile, se non grazie a una religione originata dalla funzione fabulatrice. Questa religione, che abbiamo chiamato statica, e questa obbligazione, che consiste in una pressione, sono costitutive della società chiusa». *Ivi*, p. 284.

funzione dirompente dell'intelligenza – sempre incline a soluzioni egoistiche – e, esternamente, garantisce meccanismi di identificazione o di aggressione, che la difendono contro gruppi estranei. La semplice "solidarietà sociale" in realtà racchiude "l'ostilità virtuale" tra i gruppi[36], il disprezzo latente o pubblico nei confronti dell'estraneo, giungendo a provocare persino la guerra. Solo la fraternità può forzare e rompere le barriere, ma a questa non si arriva se non attraverso un mutamento radicale dei rapporti, la cui realizzazione dipende dall'azione di alcune personalità esemplari, filosofiche e soprattutto religiose. Queste individualità privilegiate, attraverso la loro stessa vita, hanno lanciato un appello agli altri, introducendo sentimenti morali nuovi, opposti a quelli suscitati dall'obbligo sociale ordinario e, pertanto, non naturali, in quanto oltrepassano l'intenzione della natura.

Abbattuti i recinti dei gruppi, la fiducia della società chiusa viene trasfigurata[37] e trasportata su di un altro piano. Non si tratta di un semplice allargamento di confini, ma di un radicale mutamento qualitativo. La fiducia reciproca generata dall'amore veicolato dall'esempio dei modelli è assolutamente altra rispetto a quella effetto del timore e dà luogo a una morale differente in quanto all'intenzione, all'oggetto e all'azione. La fraternità non è la semplice estensione della simpatia verso il gruppo, la famiglia o la nazione tipici delle società chiuse. Jankélévitch, commentando questo aspetto de *Les deux sources*, considera un pregiudizio l'amore per le «belle gradazioni regolari», che «vorrebbe estrarre progressivamente dall'amore familiare e dal patriottismo l'amore dell'umanità»[38]. E prosegue: «Per amare l'umanità, per passare al "limite", occorre dunque una decisione improvvisa, una conversione, una "metabolé". […] Per andare dalla morale statica alla morale dinamica occorre non una "moltiplicazione", ma una conversione»[39].

La morale "aperta" presenta dunque un carattere di originalità in quanto rimanda non alle forze impersonali della natura, condensate nella pressione sociale, ma alle grandi figure morali, eroiche e uniche nella loro singolarità.

[36] *Ivi*, p. 55.

[37] *Ivi*, pp. 225-227.

[38] V. Jankélévitch, *Henri Bergson*, PUF, Paris 1959. Trad. di G. Sansonetti, *Henri Bergson*, Morcelliana, Brescia 1991, p. 235. «Ingrandire a piccole dosi la solidarietà domestica e corporativa, e, al termine di questo magnifico allargamento ottenere… la carità: che fortuna per l'egoismo! […] se nella famiglia il buon cittadino impara ad amare l'umanità o se passa continuamente dalla tribù alla patria, perché amando se stesso, non imparerebbe ad amare la sua famiglia? Al centro di tutti questi cerchi concentrici vi è evidentemente l'io che è un cerchio infinitamente piccolo, quasi un punto, in modo che la carità apparirà come il superlativo dell'egoismo!». *Ivi*, p. 236.

[39] *Ivi*, pp. 237-238.

Questi individui, a motivo della loro esemplarità, fanno appello all'imitazione da parte di altri uomini e generano un doppio movimento: quello per cui sono seguiti in quanto giudicati credibili e affidabili, dal quale origina un altro dinamismo, che introduce autentica fiducia nei rapporti sociali. L'elemento che consente questa trasformazione è la capacità di «indurre uno stato d'animo»[40], una emozione che tuttavia non appartiene né alla sfera sensibile né a quella razionale, pur partecipando di entrambe le dimensioni e che Bergson ritiene sia «di essenza metafisica ancor più che morale»[41]. I «grandi uomini di bene», in particolare i mistici cristiani, sono pertanto «rivelatori di verità metafisiche»[42]. Tuttavia il loro potere di attrazione non deriva da un'idea metafisica, bensì da quella forza che riesce a muovere la volontà, costituita per l'appunto dalla comunicazione di un'esperienza[43]. Sebbene il termine fiducia (*confiance*) non ricorra molto ne *Les deux sources*, la sequela delle personalità morali dei mistici rientra comunque in questo tipo di relazione, per la quale si decide di stringere un legame (*s'attacher*) e di conformarsi al loro esempio «come fa il discepolo con il maestro»[44].

Appare chiaramente che è il riferimento del modello a un principio che lo supera a renderne attraente e degno di fiducia l'esempio. In un altro contesto, rivolgendosi a Ferdinad Buisson, Bergson indica proprio nel rispetto della verità il motivo della fiducia che può ispirare una filosofia, in questo caso la sua:

> Avete voluto parlarmi della fiducia che i miei lavori ispirano a un certo numero di spiriti. Questa fiducia attiene principalmente al fatto che si sa che io cerco la verità prescindendo da qualsiasi retropensiero di applicazione immediata, al fatto che non sono di alcuna scuola, al fatto che non ho mai puntato a crearmene una personale e ad avere dei discepoli, al fatto che non appartengo ad alcun gruppo, infine al fatto che mi rivolgo al pubblico solo quando non posso fare altrimenti, vale a dire quando i fatti che ho raccolto e le riflessioni che essi mi suggeriscono mi portano per così dire mio malgrado a scrivere un articolo o un libro[45].

[40] Cfr. H. Bergson, *Les deux sources de la morale et de la religion*, cit., p. 57.

[41] *Ivi*, p. 248.

[42] H. Bergson, *La conscience et la vie.* Conferenza pronunciata all'università di Birmingham, il 29 maggio 1911, pubblicata con il titolo *Life and consciousness*, inclusa nel 1919 nella raccolta *L'énergie spirituelle*, in *Oeuvres*, Puf, Parigi 1959, p. 834.

[43] Bergson lo denomina uno "slancio d'amore" che si espande, capace di trasporre la vita su di un altro piano e che si trasmette a chi conforma la sua esistenza a tale modello. Cfr. *Les deux sources de la morale et de la religion*, cit., pp. 101-102.

[44] Cfr. *ivi*, p. 30.

[45] «Vous avez bien voulu me parler de la confiance que mes travaux inspirent à un certain nombre d'esprits. Cette confiance tient principalement à ce qu'on sait que je cherche la vérité en

3. *La fiducia nei modelli, principio originario di trasformazione morale*

Di poco precedenti alle riflessioni di Bergson sulla causalità esemplare dei mistici, sono le considerazioni di Scheler sul rapporto tra i valori morali e i modelli[46]. Prescindendo dalle differenze individuate dai suoi interpreti circa i distinti sviluppi della sua riflessione, è lecito affermare che il filosofo tedesco abbia dedicato una particolare attenzione al significato dell'attrazione esercitata dai modelli nella trasformazione morale, lamentando che l'etica abbia ignorato a lungo il tema, forse per il peso dell'approccio normativo. D'altra parte, egli prende le distanze dalle tesi del nominalismo etico, che vede nel genio morale non un rivelatore, ma un inventore del valore, per cui il cambiamento prodotto dalla sua persona non comporterebbe un progresso nella conoscenza etica, capace cioè di modificare l'agire, ma unicamente una nuova prassi[47].

In opposizione a questa prospettiva, per cui i valori sarebbero semplicemente dei «segni correlati ad ambiti di fatto indifferenti al valore»[48], privi dunque di validità ontologica oggettiva, Scheler ribadisce che la persona è depositaria dei valori, ma non è colei che li pone in essere[49]. La sua criti-

dehors de toute arrière-pensée d'application immédiate, à ce que je ne suis d'aucune école, à ce que je n'ai jamais visé à en créer une moi-même et à avoir des disciples, à ce que je n'appartiens à aucun groupe, enfin à ce que je ne m'adresse au public que lorsque je ne puis faire autrement, c'est à-dire lorsque les faits que j'ai recueillis et les réflexions qu'ils me suggèrent m'amènent pour ainsi dire malgré moi à écrire un article ou un livre». H. Bergson, *Lettre à Ferdinand Buisson*, 5 Juin 1912, in F. Worms (a cura di), *Annales Bergsoniennes V. Bergson et la politique: de Jaurès à aujourd'hui*, Puf, Parigi 2012, p. 42.

[46] Ci riferiamo alle considerazioni contenute soprattutto nella Parte Seconda de *Der Formalismus in der Ethik und die materiale Wertethik: Neuer Versuch der Grundlegung eines ethischen Personalismus*, M. Niemeyer, Halle a.d.S. 1913-16. Trad. di G. Caronello, *Il formalismo nell'etica e l'etica materiale dei valori. Nuovo tentativo di fondazione di un personalismo etico* (1911-1913), San Paolo, Cinisello Balsamo 1996. E agli appunti *Vorbilder und Führer* (1911-1921), raccolti da Maria Scheler nel 1933 in un'antologia dal titolo *Zur Ethik und Erkenntnislehre* e successivamente in *Gesammelte Werke*, vol. X, Francke Verlag, Berna 1957, pp. 255-354; trad. it. *Modelli e capi. Per un personalismo etico in sociologia e filosofia della storia*, a cura di E. Caminada, Franco Angeli, Milano 2011. Nonché alle riflessioni contenute nei saggi pubblicati tra il 1925 e il 1928, poi raccolti nel 1929 a cura di Maria Scheler, sotto il titolo *Philosophische Weltanschauung*, Friedrich Cohen, Bonn 1929; successivamente inseriti nel vol. IX delle *Gesammelte Werke*, a cura di M. Frings, Bouvier, Bonn 1954-1987; trad. it. *Formare l'uomo. Scritti sulla natura del sapere, la formazione, l'antropologia filosofica*, a cura di G. Mancuso, Franco Angeli, Milano 2009. Saggio introduttivo di G. Cusinato, *Rettificazione e* Bildung, pp. 7-18.

[47] M. Scheler, *Il formalismo nell'etica e l'etica materiale dei valori*, cit., p. 217.

[48] *Ivi*, p. 221.

[49] Cfr. *ivi*, p. 629.

ca si rivolge anche all'enfasi individualistica posta sui "grandi uomini", che hanno esercitato un notevole influsso sul corso della storia o sulla società, ma che spesso sono considerati tali non a motivo della perfezione ontologica, bensì per la posizione che hanno occupato in un momento storico determinato[50]. L'etica non può infatti esprimersi in un'assiologia collettivistica, ma deve attenersi al *personalismo assiologico*, ossia all'azione delle singole persone. Il "discernimento morale" o intuizione dei valori si avvale comunque in modo imprescindibile dell'autorità morale, in quanto i valori si percepiscono non soltanto attraverso un atto di introspezione o di conoscenza, ma nel "rapporto vivente, sentito" col mondo e con gli altri[51]. Egli dunque respinge qualsiasi etica autoritativa che fondi il bene e il male sul comando, mentre assegna una importanza centrale al discernimento etico e non all'obbedienza cieca. L'influenza morale di una persona su un'altra persona non sta nel suo incarnare una norma imperativa, ma «consiste nel fatto che l'intuizione pura ed immediata del suo puro valore personale e del mero essere della persona invita alla "sequela"»[52].

In questo processo intuitivo interviene il ruolo della fiducia, che Scheler qualifica come «fiducia etica», indispensabile perché l'autorità non degeneri nel suo opposto o addirittura nella violenza[53]. Per Scheler è vero che tutte le norme si fondano sui valori, ma è il valore da lui definito "personale" ad essere il più elevato (rispetto a quello reale, situazionale o legale), ossia «l'idea di una persona dotata del valore materialmente più elevato» che sarà anche «la norma suprema che regola l'esistenza e il comportamento etici»[54]. Quindi il dover-essere o ideale morale non si coglie in una norma, ma in persone che divengono modelli o ideali grazie all'intuizione dei valori elevati e positivi di cui sono dotate: «contrariamente alla norma, l'intuizione non si riferisce preliminarmente a un semplice agire, ma a un *essere*»[55]. Il modello risulta attraente in quanto una specifica obbligazione deontologica si mostra vissuta nella sua stessa vita personale. Mentre la norma è universale, sia per contenuto che per validità, il modello è connotato dall'essenza assiologico-individuale della persona, per cui

[50] Cfr. *ivi*, p. 616.
[51] Cfr. *ivi*, p. 98.
[52] *Ivi*, p. 277.
[53] «È su questo discernimento che si basa la "*fiducia*" etica nell'autorità (costituendo il fondamento essenziale della sua esistenza): ove manchi questa fiducia l'autorità scade in un potere amorale e nella violenza». *Ivi*, p. 405.
[54] *Ivi*, p. 694.
[55] *Ivi*, p. 695.

costituisce «una struttura assiologica articolata nella forma unitaria in conformità all'unità personale, una qualità di valore specifica sotto forma di persona, mentre in riferimento all'esemplarità del contenuto esso è l'unità di una istanza del dover *essere* fondata sul contenuto stesso»[56].

L'adesione fiduciosa si dà pertanto non a un ideale astratto, ma a un modello personale: è "credere in" persone, non in singole azioni[57]. È il caso, ad esempio, della forza di attrazione esercitata dal santo, alla quale ci si affida in quanto rivelatrice di un legame col divino:

> la sequela 'crede' in ciò che egli 'vede' e che essa stessa non sa 'vedere'. […] Egli è cioè il tipo di uomo che trova *fede presso la sua sequela* non perché, misurandolo con un precedente pensiero normativo, essa trova il suo agire 'buono', le sue parole 'vere', ma perché *si crede in lui*, nella sua *persona*, grazie alla qualità carismatica di questa stessa persona, del suo essere ed esser così[58].

È evidente il confronto con la posizione kantiana, che identifica il valore con la norma o la legge, un a priori di cui la persona è espressione. Per Scheler invece i modelli sono *anteriori* alle norme[59]: anzi, per comprendere le norme, occorre risalire ai modelli. È la persona stessa con i suoi atti ad essere anteriore alla norma, nel senso che, mostrandone la realizzazione nella sua vita buona, ne garantisce la positività assiologica[60]. Quindi la sequela si caratterizza come una relazione personale con il contenuto del modello «fondata sull'amore verso questo contenuto, nella formazione del suo essere *etico* personale – non quindi un'imitazione degli atti del modello o addirittura una mera ripetizione delle sue azioni o dei suoi modi di esprimersi»[61].

Il modello è fonte di *buon esempio*: in questa relazione, i valori positivi

[56] *Ivi*, pp. 701-702.

[57] «Il credere e il non-credere nell'interezza delle persone che abbiamo imparato a *comprendere* appieno non dall'esteriorità – come colui che imita solamente – ma *a partire dal loro centro di vita spirituale*, attraverso *trasposizione interiore* in esse e *compimento con* esse dei loro atti del sentire: questa è la forma più alta, più pura, più spirituale che l'efficacia del modello può assumere». M. Scheler, *Modelli e capi*, cit., p. 71.

[58] M. Scheler, *Modelli e capi*, cit., pp. 78-79.

[59] Cfr. M. Scheler, *Il formalismo nell'etica e l'etica materiale dei valori*, cit., p. 696.

[60] «Il senso più profondo di ogni atto morale non è la realizzazione di una *legge* superiore o l'attuazione di un *ordine* dotato di una struttura specifica, bensì un *regno* solidale delle *migliori persone*; in tal modo viene cioè colto nella persona non un mero soggetto di possibili atti di ragione – cioè una "persona di ragione" – bensì un centro di atti individuale, concreto, dotato in sé di valore». *Ivi*, p. 695.

[61] *Ivi*, p. 697.

da lui incarnati divengono, infatti, determinanti per favorire nell'altro lo sviluppo di valori personali analoghi.

Non vi è al mondo nulla che induca, in modo così originario, immediato e necessario, la persona a divenire buona, come il semplice discernimento, intuitivo e adeguato, di una persona buona nella *sua* bontà. Per quanto riguarda la possibilità di divenire buono, questa relazione è *assolutamente superiore a ogni altra* possibile relazione ipotizzabile come causa di un tale processo[62].

L'efficacia della relazione di buon esempio è del tutto incomparabile con quella dell'obbedienza o prescrizione, giacché in chi agisce per pura obbedienza manca sia il discernimento autonomo e immediato del valore di quanto viene comandato, sia l'intenzione di fare il bene. Solo il buon esempio può indurre quella conversione assiologica, quella configurazione morale dell'essere e del volere che persino la cosiddetta "educazione morale" non riesce a provocare, se la si intende come semplice apprendistato di azioni buone[63]. Inoltre, esso rende possibile che l'altro conservi la propria autonomia del discernimento e del volere.

In definitiva, per Scheler «il principio del modello è ovunque lo strumento *elementare* di ogni cambiamento del mondo etico»[64]. È ovvio che si possa parlare anche di contromodelli o antimodelli[65], come incarnazione di valori negativi o volgari, i quali tuttavia esercitano la stessa azione di efficacia, o anche di pseudomodelli, che emergono «dal calcolo farisaico di valere come modello meramente sul piano sociale»[66]. Saranno buoni quei modelli che incarnano quei valori più elevati ordinati gerarchicamente, quali il santo, il genio, l'eroe, il leader e l'artista[67]. Eppure, anche l'influenza dei contromodelli resta una ulteriore dimostrazione "che la persona etica può essere indotta alla conversione in modo originario (anteriore

[62] *Ivi*, p. 697.

[63] «L'anima non viene formata e plasmata da regole morali astratte e universalmente valide, ma solo e sempre da modelli chiaramente intuibili». *Modelli e capi*, cit., p. 59.

[64] *Il formalismo nell'etica e l'etica materiale dei valori*, cit., p. 698.

[65] Scheler individua in ogni società «un intero *sistema di persone sociali fungenti da modelli idealtipici* (il maestro, il vate, il poeta, il santo, il dandy), capaci di incidere *in modo originario* sia come modello, sia come antimodello su ogni processo di trasformazione morale nel bene o nel male, nel sublime o nel volgare». *Ivi*, pp. 699-700.

[66] *Ivi*, n 248, p. 703.

[67] Ci potrà essere un cattivo condottiero, ma non un cattivo santo o un cattivo eroe. *Ivi*, p. 709. Cfr. anche *Modelli e capi*, cit., pp. 72 e ss. Nell'*Osservazione conclusiva* del *Formalismo* (pp. 579-580), Scheler dichiara che è comunque indispensabile prolungare e completare la trattazione dei modelli e contromodelli con la dottrina su Dio, giacché l'etica rimanda necessariamente a una filosofia della religione.

cioè a ogni incidenza della norma o dell'educazione) solo e sempre da un'altra persona o da un'idea della medesima"[68].

L'incidenza del modello non è fondata né sulla semplice conoscenza, né sulla volontà, né sull'aspirazione né sull'imitazione o riproduzione, bensì sulla conoscenza assiologica, ossia valutativa (sentire, preferire, amare, odiare: "questo è quanto amo o odio"). Dal punto di vista della modalità di azione, il modello esercita un potere di attrazione, qualificata come irresistibile e suadente, che non è né un potere di coercizione né di suggestione, ma si fonda "su una coscienza del dover essere o della giustizia"[69].

Anche nel saggio *Le forme del sapere e la formazione*[70], trattando degli stimoli positivi che favoriscono la *Bildung*[71], intesa come processo formativo, Scheler segnala che «il primo e più importante tra essi consiste nel *modello assiologico* offerto da *una persona* che si è conquistata il nostro amore e la nostra ammirazione. *Per prima cosa*, se vuole essere "formato", l'*intero* uomo deve abbandonarsi a un modello integro e autentico, libero e nobile. Si danno anche evoluzioni in senso contrario al modello; bisogna evitarlo! Un tale modello non lo si "sceglie". È il modello che ci afferra, seducendoci e invitandoci, attirandoci impercettibilmente al suo seno»[72]. L'azione del *Vorbild*, ossia dell'esemplarità valoriale, a differenza di quella del *Führer*, capo, che massifica attraverso la manipolazione e richiede una cieca sottomissione, è invece creativa spingendo a realizzare la propria vocazione individuale: «Le personalità esemplari devono renderci *liberi* e ci rendono liberi – nella misura in cui esse stesse non sono schiave, ma libere –, liberi per l'accoglimento della nostra destinazione e per la compiuta emanazione della nostra forza»[73].

Ad attrarre non è la persona in se stessa, ma il valore "sotto forma di

[68] *Ivi*, p. 698.

[69] «I modelli attraggono verso di sé la persona che li considera tali: *non* ci si rivolge ad essi in modo attivo; il modello *diviene determinante* in riferimento al fine, pur non essendo seguito così come si persegue un fine o posto come si pone uno scopo». *Ivi*, p. 702.

[70] *Die Formen des Wissens und die Bildung*, lezione tenuta il 27 gennaio 1925 alla *Lessing Hochschule* di Berlino, poi pubblicata dall'editore Friedrich Cohen, Bonn 1925. Trad. it. *Formare l'uomo*, cit., pp. 49-89.

[71] Scheler definisce la *Bildung* come «la configurazione complessiva presa da un particolare essere umano, da non intendersi, tuttavia, al modo della forma di una statua o di un quadro […] bensì come impronta e configurazione assunte da una totalità vivente nella forma del *tempo*, di una totalità che consiste soltanto di decorsi, processi, atti». *Le forme del sapere e la formazione*, in M. Scheler, *Formare l'uomo*, cit., pp. 54-55.

[72] *Ivi*, p. 71.

[73] *Ivi*, p. 72.

persona", ossia il valore in quanto vissuto in una realizzazione individuale e come tale reso presente e percepibile nella sua esemplarità. Il modello "attrae e invita" a seguirlo, provocando attraverso la progressiva identificazione con la sua struttura e i suoi tratti[74], una trasformazione eticamente rilevante, impossibile da realizzarsi tramite il comando, il suggerimento pedagogico, il consiglio o l'orientamento. Questi ultimi, infatti, non riescono a trasformare l'intenzione etica, che comprende non soltanto il modo di pensare e di volere, ma soprattutto la conoscenza assiologico-morale, ossia il preferire, l'amare e l'odiare, che sono alla base di ogni opzione. Si tratta di «donarsi *liberamente* al contenuto assiologico personale del modello in quanto termine di discernimento autonomo»[75]. Ogni progresso assiologico-morale è pertanto regolato dall'influsso di modelli, che sono le forme più originarie dei cambiamenti etici.

4. *Conclusioni: esemplarità, fiducia e ammirazione alla luce della* Exemplarist Ethics

Le riflessioni di Bergson e Scheler sulla causalità esemplare pongono alcune questioni che meritano di essere approfondite relative alla natura della relazione di fiducia che si instaura con i modelli. L'intuizione del valore teorizzata da Scheler e il richiamo di attrazione oggetto della riflessione bergsoniana presentano senz'altro una componente cognitiva, ma ci si può chiedere fino a che punto consentano una definizione oggettiva della bontà del modello. Proporre un'etica fondata sulla forza dell'esempio e sulla testimonianza possiede indubbiamente una efficacia maggiore rispetto a un'etica rigorosamente normativa, per la sensibilità odierna e nel contesto dell'attuale crisi di autorevolezza, giacché la proposta valoriale avviene primariamente non attraverso un giudizio razionale, ma all'interno di una relazione, grazie alla mediazione pratica del rapporto io-tu. Resta da capire quanto la fiducia con la conseguente funzione performativa del modello risponda all'effettivo riconoscimento intellettuale di una eccellenza morale. Nelle riflessioni di Bergson si argomenta più volte che la causalità morale è affidata alla relazione vitale con quanto procede da un'esperienza, purché questa sia capace di svelare una verità di ordine superiore, metafisico: solo in quanto tale sarà capace di muovere positivamente la volontà.

[74] *Il formalismo nell'etica e l'etica materiale dei valori*, cit., p. 703.
[75] *Ivi*, p. 704.

Anche in Scheler, come si è evidenziato, la funzione formativa e performativa non è affidata alla persona in se stessa, ma al valore che si presenta "sotto forma di persona".

Queste considerazioni trovano una implicita eco nella recente *Exemplarist ethics*[76], proposta da Zagzebski come alternativa all'etica eudemonistica e normativa, tentando una possibile soluzione alla questione del ruolo delle emozioni nei giudizi morali. La teoria rivaluta la funzione della fiducia, in aperta critica contro il cosiddetto "egoismo epistemico", conseguenza dell'individualismo e dell'enfasi posta sull'autonomia del soggetto, che induce a non attribuire alcuna affidabilità o valore pratico o morale agli interessi e alle credenze degli altri[77].

Per la filosofa, "eccellente" significa "ammirevole", ossia degno di ammirazione, che è «l'esperienza universale con la quale identifichiamo le persone esemplari»[78] e che può costituire la fondazione per una teoria etica globale, oltre a giocare un ruolo fondamentale nell'apprendimento delle virtù morali e intellettuali. Ammirare una persona significa considerarla ammirevole e pertanto imitabile, in quanto valore da imitare[79]. Il modo in cui l'ammirazione spinge a fidarci degli altri e a riconoscerli come autorità, per la filosofa si qualifica come una pratica pre-teoretica. Sebbene, infatti, la fiducia sia la sintesi di tre dimensioni – epistemica, affettiva, comportamentale –, questo processo non parte da un giudizio, ma dalla fiducia di sé (*self-trust*), ossia del proprio sentimento di ammirazione: «io mi fido del sentimento di ammirazione, mi fido del mio modo di vedere l'oggetto della mia ammirazione. In questo caso, mi fido che la persona è ammirabile e un valore da imitare. Quando faccio così, dato che la fiducia include la credenza, io credo che essa è ammirabile e un

[76] Cfr. L. Trinkaus Zagzebski, *Exemplarist Moral Theory*, Oxford University Press, Oxford 2017. Si resta sorpresi che l'autrice, nell'argomentare la sua teoria, si riferisca solo una o due volte alle analoghe riflessioni di Scheler e Bergson.

[77] Zagzebski distingue tre forme di "egoismo epistemico" che hanno il loro correlato nell'egoismo etico: estremo, forte e debole. Il primo respinge con assoluta autoreferenzialità qualsiasi elemento di conoscenza che non provenga da se stesso; il secondo è disposto a fidarsi delle credenze altrui quando coincidano con le proprie; il terzo può ritenere affidabili quelle testimonianze che servono ai propri interessi. Cfr. L.T. Zagzebski, *Ethical and Epistemic Egoism and the Ideal of Autonomy*, in «Episteme: A Journal of Social Epistemology», 4 (3), 2007, pp. 252-263.

[78] Cfr. L.T. Zagzebski, *Exemplarist Moral Theory*, cit., p. 2.

[79] Cfr. L.T. Zagzebski, *Admiration and The Admirable*, in «Aristotelian Society Supplementary», LXXXIX (2015), n. 1, pp. 205-221. «Quando una persona non virtuosa ammira un modello, sente un'attrazione che le dà un motivo per imitarne la vita virtuosa». *Exemplarist Moral Theory*, cit., p. 169.

valore da imitare»[80]. Non occorre, secondo l'autrice, associarvi concetti descrittivi, nel senso di ricercare cosa renda buona una persona buona, ma è sufficiente identificarla come tale: chiamiamo "buone" cose e persone che rispondono a due condizioni, in quanto desiderabili, ma soprattutto in quanto ammirevoli[81].

Nella *Exemplarist ethics* si prende in considerazione non l'ammirazione nei confronti dei talenti naturali – che non hanno una connotazione morale –, ma quella verso l'eccellenza acquisita, che comprende una gamma più ampia della tradizionale classificazione delle virtù. Può peccare di ingenuità questa eccessiva fiducia nell'ammirazione: d'altra parte l'autrice ritiene che l'atteggiamento opposto – "il generale cinismo" – di chi, per non voler ammettere l'ammirazione, passa dall'invidia al risentimento, costituisca una minaccia ben più preoccupante del possibile disaccordo sugli autentici modelli[82]. Sebbene la definisca una "emozione edificante"[83] (*uplifting emotion*), Zagzebski ammette comunque la necessità che l'ammirazione nei confronti dei modelli debba essere educata attraverso la riflessione, giacché potrebbe non essere del tutto affidabile. Occorre pertanto ricorrere a delle procedure per identificare gli *exemplars* morali, costituite essenzialmente da una attenta osservazione delle loro esistenze attraverso le tradizioni e narrazioni[84], dove l'adeguatezza dell'ammirazione e di altre emozioni è verificata dalla loro resistenza nel tempo alla riflessione consapevole (*conscientious self-reflection*)[85].

Lo sforzo è dunque quello di saldare emozioni e ragione, assegnando alla fiducia nell'ammirazione un posto di primo piano nel riconoscimento della vita buona, ma anche tenendo ferma la necessità della formazione della coscienza per discernere gli esemplari autentici dalle contraffazioni. Non sono tuttavia da minimizzare le criticità di tale teoria: la circolarità tra ammirazione e riflessione appare come un circolo vizioso, posto che per fidarsi dell'ammirazione verso un modello occorre una certa "coscienziosità" o sensibilità morale che d'altronde solo grazie a un modello si può

[80] L.T. Zagzebski, *Epistemic authority*, cit., p. 89.

[81] Cfr. L.T. Zagzebski, *Exemplarist Moral Theory*, cit., p. 30. L'autrice sottolinea, infatti, che non sempre ciò che è oggetto di ammirazione è anche oggetto di desiderio, come ad esempio il martirio.

[82] Cfr. *ivi*, p. 45.

[83] *Ivi*, p. 235.

[84] «Le narrazioni e gli studi sui modelli aiutano l'immaginazione filosofica mostrandoci la varietà di vite desiderabili». *Ivi*, p. 171.

[85] Cfr. *ivi*, pp. 47-48.

conseguire[86]. È pur vero che la *Exemplarist Ethics* intende rispondere – con maggiore o minore successo – all'attuale sfiducia nelle argomentazioni razionali con una proposta non emotivista, che allo stesso tempo non richieda apriori una concezione condivisa di *télos* umano e di bene[87].

La trattazione svolta sin qui ci consente di evidenziare alcuni aspetti che non intendono comunque proporsi come vere e proprie conclusioni. Da un lato, appare chiaro che le regole astratte e la conoscenza teorica dei principi sono insufficienti a promuovere lo sviluppo morale, che avviene sempre attraverso la mediazione di una relazione personale. Gli esempi di vita buona – passati o presenti – e i modelli di eccellenza mostrano a livello pratico l'attrattiva della virtù, riuscendo a influire sugli affetti e ad agire sulla volontà. Il movimento della fiducia si genera proprio nel riconoscimento di questa esemplarità: significa credere alla praticabilità del bene, in quanto attestato da esistenze personali e rinunciare all'autosufficienza, che renderebbe impossibile attendersi dall'altro il contributo indispensabile per scoprire e giudicare se stessi.

English title: Exemplarity, an untimely proposal of meaning put to the test of trust. An analysis from Bergson and Scheler.

Abstract

The paper explores the category of exemplarity of which trust is an essential element, analysing it within its function of epistemic authority, hence in its trustworthiness and reliability, as well as a force generating trusting relationship, which create authentic social transformation, while encouraging the break of conformism and attachment to the own group. As essential need of human being, trust may, however, be declined in surrogates which produce forms of exclusive social cohesion, where reassurance is sought in expulsion of the stranger or even in war. Also, trust in an exemplar is not free from ambiguity, unless it is reported to objectivity of values witnessed by that exemplar: hence its "untopicality" within a relativistic context. The thought of Bergson on the closed society, in which trusting bonds are transfigured by

[86] Si vedano i rilievi critici mossi da C. Anderson, *Epistemic Authority and Conscientious Belief*, in «European Journal for Philosophy of Religion», 6 (2014), n. 4 pp. 91-99.

[87] È quanto afferma l'autrice nella conclusione del suo saggio, mostrando di non condividere la tesi espressa da MacIntyre in *After Virtue*. *Ivi*, pp. 230-235.

some exemplary personalities, is compared with the reflection – a few years earlier – of Scheler on the exemplars "bearers of values". Extremely actual considerations in sight of the recent "Exemplarist ethics" suggested by Zagbeski as alternative to eudaimonistic and normative ethics.

Keywords: trust; exemplarity; Bergson; Scheler; Zagbeski.

Maria Teresa Russo
Dipartimento di Scienze della Formazione - Università Roma Tre
mariateresa.russo@uniroma3.it

T

Fiducia e virtù

Giacomo Samek Lodovici

Il presente contributo si prefigge di ragionare su alcuni (non certo tutti) aspetti del rapporto tra virtù e fiducia. Quest'ultima, di per sé, non è sempre virtuosa (può infatti essere sconsiderata, irragionevole, riposta senza una minima valutazione della persona a cui viene data, ecc.), ma (come si cercherà di argomentare) può esserlo e diventarlo, può generare virtù ed essere generata da quest'ultima.

1. *Le virtù (cenni)*

In sede preliminare definiamo qui (nel solco di Aristotele)[1] le *virtù etiche* quali propensioni del soggetto a compiere atti moralmente buoni[2], che conseguono l'armonia e la sinergia tra le varie dimensioni umane (corporeità, emotività-affettività, volontà, ragione), in vista del bene complessivo della persona (il quale include una qualche soddisfazione – di cui la ragione

[1] Ci sono diverse concezioni della virtù (aristotelica, humeana, utilitarista, nietzscheana, ecc.), cfr. per es., recentemente, S. Van Hooft (ed.), *The Handbook of Virtue Ethics*, Acumen, Durham 2014; R. Hursthouse, *Virtue Ethics*, in E. Zalta (ed.), *Stanford Encyclopedia of Philosophy 2016*, https://plato.stanford.edu/entries/ethics-virtue/; D. Carr, J. Arthur, K. Kristjánsson, *Varieties of Virtue Ethics*, Palgrave-Macmillan, London 2017; A. Campodonico, M. Croce, M.S. Vaccarezza, *Etica delle virtù. Un'introduzione*, Carocci, Roma 2017.

La prospettiva che seguiamo nel presente contributo è ispirata ad Aristotele, *Etica Nicomachea* ed *Etica Eudemia*. Poiché non è possibile qui argomentare la scelta di questa aretologia, cfr. G. Samek Lodovici, *L'emozione del bene. Alcune idee sulla virtù*, Vita e Pensiero, Milano 2010.

[2] La stessa possibilità dell'esistenza della virtù etica è spesso contestata ed apre diverse questioni che richiederebbero una lunga trattazione. Qui posso solo rinviare a G. Samek Lodovici, *op. cit.*

 DOI: 10.4454/teoria.v39i1.60

deve farsi carico – appunto di tutte le dimensioni della persona, compresa quella emotiva).

Dal canto loro, le *virtù intellettuali* sono delle propensioni a compiere atti intellettivi che colgono il vero.

Correlativamente, le azioni virtuose sono quelle che promanano dalle virtù e che conseguono il bene morale (virtù etiche) e il vero (virtù intellettuali)[3].

L'effetto secondario delle virtù è facilitare, rendere spontanea e gradevole (o comunque meno difficile) l'azione buona.

Il loro effetto primario[4] (non sempre colto dagli eticisti, che colgono più spesso quello secondario) è abilitare il soggetto a comporre delle concrete azioni buone e una condotta di vita buona mediante il desiderio-proposito (tramite le virtù etiche), l'individuazione nella situazione concreta (grazie alla virtù intellettuale della *phronesis*) e la scelta-esecuzione (di nuovo grazie alle virtù etiche) dell'azione buona in una situazione particolare: senza virtù è frequentemente[5] impossibile il desiderio-proposito e l'individuazione-gestazione-esecuzione della concreta azione buona.

La virtù è dunque un potenziamento dei principi operativi umani:

– potenzia la ragione nell'individuazione-valutazione-comando dell'azione virtuosa;
– potenzia e coltiva le emozioni, inclinandole sia a fornire la propria energia alle azioni virtuose, sia a percepire gli aspetti eticamente salienti di una situazione concreta;
– potenzia la volontà[6] nel desiderio di compiere azioni virtuose e nell'esecuzione della scelta (però il virtuoso desidera comportarsi moralmente bene, ma non è necessitato ad agire dalla virtù).

[3] Il coglimento del vero sul bene non è sufficiente per volerlo realizzare (vs intellettualismo etico); d'altro canto, lo stesso esercizio di atti intellettivi che raggiungono la verità è quasi sempre moralmente buono, cfr. G. Samek Lodovici, *op. cit*., pp. 147-150 e L. Zagzebski, *Virtues of the mind. An inquiry into the nature of virtue and the ethical foundation of knowledge*, Cambridge University Press, Cambridge 1996.

[4] Cfr. G. Samek Lodovici, *op. cit.*, pp. 9-15, 80, 108, 122, 181-195.

[5] Cfr. *ivi*, capitolo VI.

[6] Qui accenniamo, storiograficamente, che Aristotele menziona le volizioni, ma la presenza di una nozione di volontà nel suo pensiero è da alcuni negata: non vogliamo entrare in questa controversia interpretativa.

La volontà ha acquisito una rilevanza molto forte in diverse antropologie filosofiche e teologiche specialmente a partire da Agostino (che pur non ne è certo lo scopritore).

2. *Fiducia e af-fidamento*

Qui di seguito intenderemo l'atto di fede-fiducia[7] di un *trustor* verso un *trustee* in due accezioni: pratica ed epistemologica.

I. In senso pratico, qui di seguito, la fiducia designa l'attribuzione[8] speranzosa di affidabilità al *trustee* (compreso il sé, nel caso della fiducia in se stessi) ed al suo agire nella speranza (talvolta molto solida e quasi certa, talvolta più esile) di una sua positiva risposta – tramite un atteggiamento/reazione/comportamento positivo (positivo in senso pre-morale) – alla fiducia che gli abbiamo concesso.

Di conseguenza la fiducia comporta la formulazione di richieste di aiuto e/o il conferimento all'altro di una serie di compiti, mansioni, e/o di libertà di scelta e di libertà d'azione, l'assegnazione all'altro di uno spazio d'azione, e/o l'assegnazione di un utilizzo delle cose (anche di grande valore) appartenenti al *trustor*.

La fiducia è un'attribuzione di affidabilità talvolta fondata sulla credenza (corretta o errata) che questa affidabilità sia già presente nel *trustee*, talvolta ritenendo (correttamente o erroneamente) che essa non sia ancora sussistente ma possibile (come vedremo al § 4).

II. In senso più epistemologico, qui di seguito, la fiducia designa l'attribuzione di affidabilità all'attestazione del *trustee*, alle sue affermazioni.

Quando il *trustee* è destinatario di entrambe queste due forme di fiducia, la fiducia che gli esprime il *trustor* è l'attribuzione di affidabilità globale da parte di un soggetto ad un altro soggetto («mi fido di te in tutto»)[9].

In più, probabilmente si può precisare con Annette Baier che la fiducia include sì un affidamento, ma che non ogni affidamento è fiducia, bensì solo quello che con-fida nell'esistenza di una (almeno minima) buona volontà del *trustee* nei confronti del *trustor*. Per esempio, un partner commerciale, e perfino un soggetto che so essere malintenzionato nei miei confronti, posso ritenerli affidabili se reputo (a torto o a ragione) che sia per loro utile rispettare un accordo o un contratto con me[10], senza però che io nutra vera

[7] Una rassegna di diverse teorie filosofiche recenti sulla fiducia in O. Lagerspetz, *Trust, Ethics and Human Reason*, Bloomsbury, London-New York 2015.

[8] Tale atto può essere accompagnato da sentimenti di fiducia, serenità, tranquillità (di cui non mi occuperò in questo contributo), ma non coincide con essi.

[9] Luhmann ritiene di rilevare delle differenze tra confidare e fiducia in N. Luhmann, *Familiarity, Confidence, Trust: Problems and Alternatives*, in D. Gambetta (ed.), *Trust. Making and Breaking Cooperative Relations*, Basil Blackwell, New York 1988, pp. 95-106.

[10] A. Baier, *Trust and Antitrust*, in «Ethics», 96 (1986), 2, pp. 231-260, pp. 234-235. Cfr.

fiducia in loro. Del resto, già Platone[11] notava che persino in una banda di delinquenti si fa affidamento reciproco, altrimenti i delinquenti non potrebbero sensatamente cominciare a concretizzare dei progetti, perché non potrebbero mettersi a cooperare, e la banda si scioglierebbe.

Inoltre, senza peraltro voler tracciare differenze troppo nette, si può probabilmente convenire con Hertzberg[12], quando dice che si fa affidamento su un'altrui attività circoscritta, mentre la fiducia investe più globalmente la persona che esercita la tal o talaltra attività.

Ad ogni modo, qui di seguito a volte parleremo di fiducia senza distinguerla dall'affidamento.

Quanto all'atteggiamento verso le attestazioni altrui, noi esplichiamo atti di affidamento e (se presupponiamo anche una minima volontà buona nei nostri confronti) anche atti di fede-fiducia non solo se crediamo in Dio[13], se crediamo a chi ce ne parla, ma altresì ogni volta che riteniamo vero un X di cui non abbiamo fatto esperienza diretta e che apprendiamo dai vari mass media, dai libri, dal web, ecc., nonché da amici, parenti, colleghi, ecc.

I latini dicevano che almeno *mater est semper certa*, ma Agostino notava[14] che non possiamo essere certi nemmeno dell'identità di nostra madre[15] perché quando siamo stati partoriti eravamo inconsapevoli: lo sappiamo per fede. È vero che i genitori solitamente somigliano ai figli, ma esistono casi di somiglianza anche tra persone che non sono consanguinee ed esistono anche i sosia, perciò noi crediamo per fede che i nostri genitori siano davvero quelli che pensiamo. E, se cerchiamo di accertare la nostra origine col test del DNA, dobbiamo af-fidarci ai tecnici che svolgono il test. Insomma, l'essere umano fa continuamente atti di af-fidamento e di fede-fiducia.

anche F. Ruokonen, *Trust, Trustworthiness, and Responsibility*, in P. Mäkelä, C. Townley (eds.), *Trust: Analytic and Applied Perspective*, Rodopi, Amsterdam-New York 2013, pp. 1-14, p. 5. In questo articolo la Baier critica la frequente focalizzazione del contrattualismo (spesso professato da filosofi maschi) appunto sulle relazioni contrattuali simmetriche, a scapito delle relazioni non contrattuali (e spesso asimmetriche), per esempio appunto quelle fiduciarie, molto più valorizzate dalle donne filosofe (o psicologhe, ecc.), *ivi*, pp. 246-253.

[11] Platone, *Repubblica*, 351 C - 353 C. Cfr anche Agostino, *De civitate Dei*, IV, 4.

[12] L. Hertzberg, *On The Attitude of Trust*, in «Inquiry», 31 (1988), pp. 307-322, 312.

[13] È la fede come virtù teologale, di cui non mi occupo.

[14] Agostino, *Confessiones*, 6, 5, 7.

[15] A maggior ragione oggi per i nati da fecondazione artificiale eterologa con produttrici anonime di ovociti.

3. *La necessità della fiducia*

Ora, Confucio insegnava[16] che i Capi di Stato hanno specialmente bisogno di tre cose: armi, cibo e fiducia, che è la più importante. Infatti, le armi non salvano i soldati quando questi perdono totalmente la fiducia nell'esito della battaglia e nel generale, e la fame (purché non sia estrema) non fa, da sola, crollare una nazione se questa ha fiducia nel governo che la conduce.

Ma la fiducia non è necessaria solo in uno Stato già esistente, bensì per lo stesso cominciamento della vita associata, perciò l'ipotesi hobbesiana è aporetica. Infatti, per Hobbes, gli uomini non possono riporre alcuna fiducia negli altri: è un'affermazione che discende dalla sua antropologia egoista: se noi tutti siamo *sempre* e *soltanto* alla ricerca del nostro vantaggio, nessuno può riporre la propria fiducia negli altri, perché, in una situazione pre-statale di penuria di beni, l'altro ci aggredirà sicuramente se per lui è vantaggioso farlo, perciò è insensato fidarsi. Per Hobbes gli esseri umani possono fidarsi reciprocamente solo dopo aver costituito lo Stato che fa coercitivamente rispettare le leggi e i contratti[17].

Tuttavia, si può replicare a Hobbes che, già prima della stipulazione del patto sociale che costituisce lo Stato, è necessaria proprio la fiducia (o almeno l'affidamento) per accordarsi sulle regole del patto sociale stesso, è necessario un minimo di af-fidamento/fiducia naturale pre-statale nel fatto che gli altri non mi stiano frodando e manipolando durante la stipulazione delle regole dell'interazione sociale e durante l'istituzione dello Stato.

Inoltre, come dice la O'Neill, le procedure e «i sistemi per indurre le persone a rispettare gli impegni e non tradire le attese sono anch'essi, alla fine, qualcosa di cui dobbiamo… fidarci. Viene sempre il momento in cui la fiducia non ha alternative». In effetti, «la vecchia domanda "chi sorveglierà i sorveglianti?" non ammette risposte definitive. È perché le garanzie sono sempre imperfette che la fiducia è così importante. Ogni garanzia rinvia a un ultimo garante di cui ci dobbiamo fidare»[18].

Nemmeno la divisione dei poteri è un meccanismo sufficiente, perché gli esponenti di uno dei tre poteri possono allearsi con gli esponenti degli altri o comunque possono riuscire a prevaricare.

[16] Lo riferisce O. O'Neill, *A Question of Trust*, Cambridge University Press, Cambridge 2002, trad. it. di S. Galli, *Una questione di fiducia*, Vita e Pensiero, Milano 2003, p. 35.

[17] Per una valorizzazione, in antitesi a Hobbes, della filosofia politica di Leibniz cfr. L. Pierfranceschi, *Relazionalità e fiducia. Per un'etica dei legami*, Cortina, Milano 2016, pp. 53-69.

[18] O'Neill, *cit*., p. 37.

La fiducia, poi, è una cruciale componente costitutiva della vita sociale, anche economica[19], e delle dinamiche di tutte le istituzioni[20].

Dunque, con buone ragioni nell'antica Roma (secondo la tradizione) alla radice della convivenza civile c'era il culto della dea *Fides*, a cui forse è stato dedicato il primo tempio[21].

E, del resto, anche una delle ipotesi etimologiche è utile al riguardo, visto che la radice *fid* delle lingue neolatine rimonta sia al greco *peith*, cioè legare (e persuadere), che concorre alla formazione della parola *pistis* (fede), sia ai termini sanscriti *bandh*, che significa legame[22], e la fiducia crea appunto legami sociali, e *bheidh*, che probabilmente designava[23] le parti intrecciate di una pianta (ad esempio un giunco), il che allude al legame-impegno interpersonale.

In effetti, senza affidamento/fiducia la nostra esistenza associata, a tutti i livelli delle interazioni sociali (dirette o indirette, personali o anonime,

[19] «Il momento più acuto della crisi [economica], nel quale la fiducia tra [e verso] le banche scende ai minimi storici paralizzando il mercato interbancario e i risparmiatori di alcune banche perdono la fiducia andando a ritirare i loro risparmi agli sportelli, ci fa capire che senza fiducia (nelle banche, nella moneta) il sistema non può sopravvivere […]. Tutte le relazioni interpersonali in economia si realizzano in un contesto di informazioni incomplete (non sappiamo mai fino in fondo chi abbiamo davanti) e di incompletezza dei contratti (non possiamo garantirci da comportamenti scorretti calcolando e contrattualizzando tutte le evenienze possibili). Pertanto se la fiducia è più elevata siamo più produttivi», soprattutto oggigiorno: «Le attività sempre più complesse sviluppate nelle imprese richiedono un lavoro di gruppo nel quale competenze non sovrapponibili di esperti di settore (diritto, economia, tecnologia, marketing) devono poter interagire tra loro per poter ottenere un risultato soddisfacente. Senza atteggiamento pro-sociale e la fiducia che induce i partecipanti a condividere quello che sanno con gli altri senza temere il rischio di essere abusati per questo l'impresa non è in grado di raggiungere livelli di produttività soddisfacenti», L. Becchetti, *I vizi, le virtù e il mercato*, http://www.benecomune.net/archivio/focus/i-vizi-le-virtu-e-il-mercato-2/.

La fiducia, già importante in un'economia del baratto, diventa necessaria in un'economia che si basa sul denaro: ci vuole fiducia nel potere centrale che emette la moneta, fiducia che la persona che ci paga una cosa non stia spacciando valuta falsa, fiducia che il denaro accettato ora possa venire speso successivamente con il medesimo valore (o almeno a un valore poco inferiore), ecc., G. Simmel, *The Philosophy of Money*, Routledge, London-New York 1990, p. 178. L'iscrizione «*non aes sed fides*», che si trova sulle monete di Malta, indica il ruolo centrale della fiducia (*ibidem*). Cfr. anche R. Grandini, *La dimensione "religioide" della fiducia istituzionale e interpersonale nella sociologia di George Simmel*, in B. Cattarinussi (a cura di), *Emozioni e sentimenti nella vita sociale*, FrancoAngeli, Milano 2000, pp. 201-226.

[20] Per quanto segue nel presente paragrafo sono specialmente debitore verso V. Pelligra, *I paradossi della fiducia. Scelte razionali e dinamiche interpersonali*, il Mulino, Bologna 2007, da cui traggo anche alcune citazioni.

[21] Almeno secondo G. Freyburger, *Fides. Étude sémantique et religieuse depuis les origines jusqu'à l'époque augustéenne*, Les Belles Lettres, Paris 1986, pp. 259-262.

[22] S. Bittasi, *Fiducia*, in «Aggiornamenti sociali», 5 (2010), pp. 391-394, p. 391.

[23] L. Pierfranceschi, *op. cit.*, pp. 18-19.

ecc.), diventerebbe invivibile, in quanto gran parte della nostra vita dipende dall'agire altrui[24], su cui abbiamo bisogno di fare affidamento[25].

Come dice Mill, «il vantaggio per l'umanità di essere capace di fidarsi reciprocamente pervade ogni spiraglio e ogni fessura della vita umana»[26]. Non solo, come dice Arrow, la fiducia è «il lubrificante del sistema sociale» e «potersi fidarsi risparmia una enorme quantità di problemi»[27], ma, inoltre, senza af-fidamento la cooperazione scomparirebbe quasi del tutto e (come dice Luhmann)[28] non potremmo quasi alzarci dal letto la mattina.

4. *La fiducia genera affidabilità e virtù*

Ora, se consideriamo la fiducia nella sua accezione pratica (nel senso I menzionato al § 2), quale attribuzione speranzosa di affidabilità al *trustee* ed al suo agire, nella speranza di una sua positiva risposta, è vero che, *a volte, «non fidarsi è meglio»*, tuttavia spesso la fiducia genera affidabilità, cioè la mia affidabilità è generata spesso dalla tua fiducia: il comportamento del *trustor* produce non infrequentemente l'affidabilità del *trustee*[29], perché un esplicito atto di fiducia, tanto più se virtuoso (preciserò meglio via via che cosa intendo per atto virtuoso di fiducia), può indurre, non di rado, un'affidabile risposta virtuosa o quasi-virtuosa.

E quando un atto di fiducia riceve come risposta un tradimento, un ulteriore atto di fiducia (verso lo stesso soggetto), tanto più se virtuoso, può generare (anche se, ovviamente, ciò non accade sempre), almeno questa volta, una risposta virtuosa o quasi. Un po' come succede ne *I Miserabili* di Victor Hugo: il vescovo Myriel accoglie fiduciosamente in casa lo sbandato Valjean, che è stato in carcere per 19 anni; questi, però, ruba l'argenteria e fugge. Valjean viene poi arrestato e ricondotto da Myriel, il quale lo scagiona, raccontando ai gendarmi di aver donato a Valjean l'argenteria e garantendo sulla sua futura buona condotta; dopodiché si rivolge al ladro

[24] Per es., se esco di casa e prendo l'auto vuol dire che, almeno in una certa misura, faccio affidamento sul rispetto del codice stradale da parte degli altri, sull'esistenza di stazioni di rifornimento, sulla tenuta dei ponti che devo percorrere, ecc.

[25] Cfr. anche C. Townley, J.L. Garfield, *Public Trust*, in P. Mäkelä, C. Townley (eds.), *op. cit.*, pp. 95-107.

[26] J.S. Mill, *Principles of Political Economy with Some of Their Application to Social Philosophy*, University of Toronto Press-Routledge and & Kegan Paul, Toronto-London 1965, p. 110.

[27] K. Arrow, *The Limits of Organization*, Norton, New York 1974, p. 23.

[28] N. Luhmann, *Trust and Power*, Wiley, Chichester 1979, p. 4.

[29] V. Pelligra, *I paradossi della fiducia*, specialmente pp. 169-178, pp. 237-257.

dicendogli di aver molta fiducia in lui, nella sua capacità e volontà di diventare onesto: il che effettivamente accade, perché il comportamento di Myriel genera in Valjean un virtuoso desiderio-proposito virtuoso (cfr. § 1) di cambiar vita[30].

Quanto detto fa emergere anche che la fiducia virtuosa è alimentata da una speranza *ragionevole* (altrimenti accordare la fiducia è vizioso), fa emergere la connessione tra la virtù della fiducia e quella della speranza, che è[31] un'attesa fiduciosa di un bene possibile, ma che non è con certezza raggiungibile né a portata di mano, che è almeno parzialmente arduo, e il cui raggiungimento dipende solo in parte, o a volte per nulla, dal soggetto.

In altri termini, la fiducia altrui può generare una risposta virtuosa o quasi-virtuosa di un soggetto:

- a volte perché vogliamo ottenere l'approvazione-riconoscimento-apprezzamento degli altri per tale nostra risposta;
- a volte perché vogliamo evitare l'afflizione che ci provoca la disapprovazione e/o il risentimento altrui (che sono maggiori se l'attestazione di fiducia che abbiamo ricevuto è stata fatta davanti a molte altre persone) per una nostra risposta contraria alle aspettative;
- a volte perché genera in noi la percezione di un'obbligazione («il solo fatto che qualcuno abbia riposto la sua fiducia in noi ci fa sentire obbligati, e questo rende più difficile tradirla»)[32], che ci fa desiderare di non tradirla;
- a volte[33] perché genera in noi un senso di autostima, che ci sospinge a desiderare di essere virtuosi per amore dell'azione virtuosa in se stessa (su questo tornerò fra poco).

D'altra parte, ancorché più raramente, anche la fiducia come affidamento all'attestazione altrui, poiché manifesta al *trustee* che il *trustor* lo

[30] Un altro esempio letterario è quello manzoniano della conversione dell'Innominato. Quando il «selvaggio signore», che si è macchiato di molti delitti, incontra Lucia, la forza interiore e il cuore virtuoso di quest'ultima, le sue parole, la sua capacità di riconoscere le tracce di bontà d'animo dell'altro, e la sua fiducia non solo nella Provvidenza, ma anche nella capacità dell'Innominato di operare una *metanoia*, risvegliano la capacità di pentimento e di bene di quel potente. Lucia dice all'Innominato: «Vedo che lei ha buon cuore, e che sente pietà di questa povera creatura […]. Compisca l'opera di misericordia: mi liberi, mi liberi», A. Manzoni, *I Promessi sposi*, Le Monnier, Firenze 1986, p. 365.

[31] Cfr. J. Pieper, *Hoffnung und Geschichte*, Kösel-Verlag, Munchen 1967, trad. it. *Speranza e storia*, Morcelliana, Brescia 1967, pp. 14-23.

[32] P. Dasgupta, *Trust as a Commodity*, in Gambetta (ed.), *Trust*, cit., p. 53.

[33] Lo aggiungiamo noi alla disamina di Pelligra.

considera credibile, sincero, informato e competente su quanto il *trustee* appunto attesta, può anzitutto generare o rinforzare l'autostima nel *trustee*, e può generare le sue risposte virtuose o quasi-virtuose per i motivi appena visti, specialmente per il desiderio di confermare la qualità-patente di affidabilità che gli è stata attribuita.

Ovviamente gli atti fiduciali non ottengono sempre risposte positive e ribadiamo che «a volte non fidarsi è meglio» ed a volte è giusto e necessario punire. Ma «esiste ormai un'abbondante letteratura empirica che mostra, attraverso studi sul campo e di laboratorio, l'esistenza di una spiccata propensione a fidarsi degli altri e a ripagare tale fiducia», cioè «si è venuta ad accumulare un'ampia evidenza empirica che mostra come i soggetti reali tendano a fidarsi e a ripagare la fiducia in proporzione molto maggiore di quanto non prevede la teoria standard»[34], quella che si basa sul concetto di *homo oeconomicus* esclusivamente autointeressato.

Dunque, «se ci comportiamo come se ci aspettassimo il meglio dagli altri, essi, spesso, come risultato si comporteranno meglio»[35], perché non di rado «le aspettative [comunicate] di una persona circa il comportamento di una seconda conducono quest'ultima ad agire in modo da confermare le aspettative originali della prima»[36].

Inoltre, non sono soltanto gli altri a motivarci ad agire bene: non soltanto noi desideriamo essere approvati, ma desideriamo anche la nostra autoapprovazione e desideriamo anche essere degni di stima e di lode. Come dice Smith, «L'uomo desidera naturalmente, non solo di essere amato, ma anche di essere amabile [...]. Egli teme, naturalmente, non solo di essere odiato, ma anche di essere odioso [...]. Egli desidera non solo la lode, ma anche di essere degno di lode»[37], ecc.

Insomma, la risposta positiva del *trustee* viene provocata dall'utilità e/o dal desiderio di approvazione (di stima) e/o dall'avversione per la disapprovazione (per la disistima) e/o dal desiderio di autoapprovazione (di autostima) e/o dall'avversione per l'autodisapprovazione (per l'autodisistima)[38] e/o

[34] V. Pelligra, *Fiducia*, in L. Bruni, S. Zamagni (a cura di), *Dizionario di Economia civile*, Città Nuova, Roma 2009, pp. 397 e 401.

[35] J. Baron, *Trust: beliefs and morality*, in A. Ben-Ner, L. Putternam (eds.), *Economics, Values and Organizations*, Cambridge University Press, Cambridge 1998, p. 411.

[36] L. Jussim, *Self-fulfilling Prophecies. A Theoretical and Integrative Review*, in «Psychological Review», 93 (1986), p. 429.

[37] A. Smith, *The Theory of Moral Sentiments*, Clarendon Press, Oxford 1976, pp. 113-114.

[38] Aristotele, *Etica Nicomachea*, 1166b 15-20, rilevava che i malvagi, purché la loro coscienza morale non si sia offuscata, «Non avendo nulla di amabile, non provano alcun sentimento amorevole verso se stessi».

dall'apprezzamento-desiderio del valore intrinseco dell'azione (che reciproca la fiducia ricevuta), la quale azione, in quest'ultimo caso, è pienamente virtuosa.

E come «la gente è [più facilmente] altruista verso gli altruisti»[39], così la fiducia, tanto più se virtuosa, genera azioni virtuose o, almeno, pre-virtuose. Quando una certa richiesta viene formulata in un contesto che lascia libertà di risposta ed esprime fiducia nella cor-rispondenza positiva dei suoi destinatari, generalmente questi ultimi sono maggiormente propensi a cor-rispondere al comportamento atteso.

Un esempio di questo effetto si vede nella pratica delle donazioni volontarie. Alcuni esperimenti sociali[40] mostrano che quando gli inviti a fare donazioni vengono espressi attraverso formule come «noi contiamo su di lei», le donazioni aumentano, perché esprimono fiducia[41] verso il potenziale donatore.

Un altro esempio analogo è quello relativo al confronto tra la raccolta del sangue fatta in Inghilterra, dove avviene grazie alle donazioni, e negli Stati Uniti, dove invece è remunerata. Ebbene, il sistema di raccolta inglese ottiene (percentualmente, al netto della differenza di popolazione) più sangue e più regolarmente[42].

Ancora, un altro studio ha analizzato un altro caso[43], il comportamento di due gruppi di professionisti sottoposti a controlli differenti. Al primo gruppo era stata concessa fiducia e perciò un ampio margine di discre-

[39] M. Rabin, *Incorporating Fairness into Game Theory and Economics*, in «The American Economic Review» 83 (1993), pp. 1281-1302, p. 1281.

[40] Citati da V. Pelligra, *I paradossi della fiducia*, cit., pp. 238-240.

[41] Si tratta di una fiducia che: a) può avere solo un obiettivo pragmatico (una maggior raccolta di denaro per beneficienza) e che in tal caso non è virtuosa (senza essere per forza cattiva), perlomeno non è virtuosa nei confronti del *trustee*, che cerca di influenzare nella direzione voluta (ancorché i promotori della raccolta di fondi possano essere virtuosi nei confronti dei potenziali beneficiari delle somme raccolte per beneficienza); b) oppure (più raramente) può avere (anche) una virtuosa intenzione di favorire nel *trustee* un atteggiamento e un comportamento benevolente verso altri, che si esprima appunto nella donazione per beneficienza.

[42] Anche in questo esempio, la fiducia espressa da chi fa la raccolta del sangue: a) può avere solo un obiettivo pragmatico (una maggior raccolta di sangue), nel qual caso non è virtuosa (senza essere per forza cattiva), perlomeno non nei confronti del *trustee*, che cerca di influenzare nella direzione voluta (ancorché i promotori della raccolta del sangue possano essere virtuosi nei confronti dei potenziali beneficiari del sangue raccolto); b) oppure (più raramente) può avere (anche) una virtuosa intenzione di favorire nel *trustee* un atteggiamento e un comportamento benevolente verso altri, che si esprima nella donazione del sangue.

[43] Anche in questo caso, così come nel prossimo esempio che citiamo subito dopo, possono verificarsi le due fattispecie di fiducia menzionate nelle precedenti note a proposito della beneficienza e della raccolta del sangue.

zionalità, mentre il secondo gruppo veniva controllato in maniera pressante: la produttività più alta era quella del gruppo a cui era stata concessa fiducia[44].

Similmente[45], la Graamen Bank del Bangladesh, la quale concede prestiti sulla fiducia a soggetti impossibilitati a fornire solide garanzie di restituzione, riceve la restituzione dei prestiti nel 94% dei casi, in un paese dove invece il sistema bancario quando chiede solide garanzie sperimenta tassi di recupero bassissimi.

Dal canto suo, Mark Alfano, sulla scorta di diversi riscontri empirici ricavati da varie ricerche sociali, ha focalizzato[46], svolgendo anche argomentazioni simili alle nostre[47], quelle che lui chiama *factitious virtues*, cioè quelle virtù che, in un soggetto X, si ingenerano o si consolidano quando un soggetto Y, in forza di un atteggiamento fiduciale (tralasciato da Alfano) attribuisce tali virtù appunto al soggetto X: delle virtù che, nel momento dell'attribuzione, il soggetto X non possiede effettivamente o che possiede solo parzialmente, ma che l'attribuzione fatta da Y fa insorgere o consolida[48].

Questi ed altri esempi, tralasciando le loro differenze, manifestano un tratto comune, cioè la «presenza di una struttura di interazione di tipo fiduciario» e «un modello di comportamento che appare essere radicalmente in contrasto con le previsioni del modello economico tradizionale».

[44] H. Barkema, *Do Top Manager Work Harder When They Are Monitored?*, in «Kyklos», 48 (1995), pp. 19-42.

[45] Lo riferisce sempre V. Pelligra, *I paradossi della fiducia*, cit., p. 240.

[46] M. Alfano, *Character as Moral Fiction*, Cambridge University Press, Cambridge 2013, specialmente pp. 82-108.

[47] Aggiungendo dei paragoni con le conseguenze reali di un placebo e con le conseguenze reali delle profezie che si autoavverano.

[48] Alcune ricerche sociologiche che manifestano l'esistenza delle *factitious virtues* sono riferite da M. Alfano, *op. cit*., alle pp. 89-90. Ad esempio, una ricerca riferita da Alfano riguarda un docente che, facendo, per un certo numero di giorni, una serie di elogi ad una classe di studenti, attribuendo loro la virtù dell'ordine ed una già effettiva premurosa cura della pulizia dell'aula, è riuscito effettivamente a promuovere questi comportamenti, che prima non sussistevano. Un significativo miglioramento del comportamento in senso ecologico è stato rilevato presso un gruppo di adulti che venivano elogiati come già caratterizzati da tale sensibilità e condotta; un miglioramento decisamente maggiore rispetto a quello prodotto illustrando gli effetti nocivi e dannosi sull'ambiente di certe condotte. In un altro studio riferito da Alfano è risultato che elogiando dei bambini dicendolo loro che erano generosi, anche se non lo erano ancora, essi diventavano più generosi di altri che erano meramente esortati a farlo. Un altro caso riportato è quello di un messaggio televisivo che elogiava il popolo americano come già desideroso di risolvere i problemi energetici e già impegnato a farlo, e che ha significativamente intensificato i comportamenti in tal senso.

Quest'ultimo «assume che le scelte dei soggetti siano influenzabili attraverso l'uso di incentivi e disincentivi materiali, i quali, facendo leva sulle motivazioni autointeressate degli agenti stessi, possono essere utilizzati per favorire o scoraggiare determinate azioni»: il datore di lavoro può pensare di ottenere che il dipendente agisca in un modo invece che in un altro attraverso incentivi e soprattutto punizioni; le istituzioni sanitarie possono cercare di favorire la raccolta del sangue retribuendo i "donatori"; le banche possono pensare di minimizzare le perdite dei prestiti richiedendo solide garanzie ecc. «Eppure i quattro esempi che abbiamo appena discusso, e molti altri[49] che potrebbero essere ricordati, mostrano come, in particolari circostanze, gli stessi incentivi materiali possano operare in direzione opposta rispetto a quanto generalmente creduto e voluto. Gli incentivi materiali invece di favorire una certa condotta [che si vuole produrre] la possono scoraggiare»[50], ed i disincentivi e le sanzioni possono produrre proprio la condotta che volevano scongiurare[51].

Ovviamente, le cause di questi dati di fatto sono molteplici, ma nella loro produzione incide molto ciò che abbiamo sopra menzionato e cioè i desideri e le avversioni sopra messi in luce, nonché il desiderio di ripagare la fiducia altrui: un soggetto che si comporta in maniera fiduciosa con ciò stesso comunica ai suoi interlocutori che crede che essi si comporteranno positivamente e proprio ciò può motivarli a confermare le aspettative[52], il che (talvolta) in un circolo virtuoso aumenta la quantità circolante di fiducia iniziale. Di conseguenza, «la fiducia non è una risorsa che si esaurisce con l'uso: al contrario, più c'è, più è probabile che essa aumenti»[53], sia

[49] Cfr., per es., N. Guéguen, A. Pascual, *Evocation of Freedom and Compliance: The "But You Are Free of…" Technique*, in «Current Research in Social Psychology», 5 (2000), pp. 264-270.

[50] V. Pelligra, *I paradossi della fiducia*, cit., p. 241.

[51] Diverso e benefico è invece l'effetto dei premi, sulla cui differenza con gli incentivi, cfr. L. Bruni, *Le nuove virtù del mercato nell'era dei beni comuni*, Città Nuova, Roma 2012, pp. 133-139 e S. Zamagni, *Perché ritornare a Giacinto Dragonetti*, Prefazione a G. Dragonetti, *Trattato delle virtù e dei premi*, Carocci, Roma 2012, pp. 18-21. Per esempio, un incentivo monetario viene promesso prima del compimento di una prestazione (mentre un premio viene conferito a posteriori e senza essere stato promesso) e, specialmente *se reiterato nel tempo* e dato al lavoratore per svolgere ciò che fa già parte dei suoi obblighi (il caso degli straordinari è diverso), può fargli pensare che il suo datore di lavoro mediante l'incentivo gli paghi un sovrapprezzo perché non si fida del lavoratore dal punto di vista tecnico e/o morale, cosicché l'incentivo ha l'effetto di indebolire l'autostima del lavoratore. Oppure può fargli ritenere che la sua retribuzione sia ingiusta e che l'incentivo sia una parziale compensazione (non dichiarata) per una retribuzione che il datore di lavoro sa essere appunto ingiusta.

[52] V. Pelligra, *I paradossi della fiducia*, cit., p. 246.

[53] D. Gambetta, *Can we Trust Trust?* in Id. (ed.), *Trust*, cit., pp. 213-237, p. 234.

perché appunto la fiducia promuove i comportamenti positivi che la ripagano e la incrementano, sia perché i comportamenti umani, comprese (anzi forse specialmente) le azioni virtuose (compresa quell'azione virtuosa che è la concessione della propria fiducia ragionevole e bene-volente), si imparano e si diffondono specialmente attraverso l'imitazione[54].

5. *La nocività della s-fiducia*

Chi, al contrario, promette incentivi materiali e, soprattutto, minaccia sanzioni, dimostra la sua diffidenza e comunica il seguente messaggio: «penso che tu non risponderai positivamente senza remunerazioni, controlli e sanzioni», e tale messaggio è un atto di disistima che ha un effetto demotivante sul destinatario.

Insomma, molti esseri umani agiscono sulla base di motivi immateriali (approvazione altrui, autostima, valore intrinseco di un atto, desiderio di ripagare la fiducia, ecc.) e non solo sulla base dell'utilità materiale che si aspettano dalle proprie azioni.

Pertanto[55], un ambiente in cui le sanzioni e i controlli sono molto pervasivi e congegnati presupponendo che i soggetti siano strutturalmente degli opportunisti:

- può segnalare ad un soggetto potenzialmente affidabile che i comportamenti dominanti in quel luogo sono quelli opportunisti, e ciò può ingenerare un comportamento analogo anche in questo soggetto;
- può allontanare i soggetti che agiscono sulla base di motivazioni intrinseche, le quali promuovono comportamenti più efficaci, e può attrarre coloro che agiscono invece per motivazioni opportunistiche;
- può determinare per reazione comportamenti di gruppo reticenti e di omertà[56];
- può (lo ripetiamo) produrre, alla lunga, l'aumento delle trasgressioni. Infatti, come hanno evidenziato in particolare Frey[57] e Pettit[58], quando il

[54] Ho argomentato a lungo questa tesi in G. Samek Lodovici, *L'emozione del bene*, cit.

[55] V. Pelligra, *I paradossi della fiducia*, cit., pp. 251-252.

[56] P. Pettit, *Republicanism. A Theory of Freedom and Government*, Clarendon Press-Oxford University Press, New York 1997, trad. it. di P. Costa, *Il repubblicanesimo. Una teoria della libertà e del governo*, Feltrinelli, Milano 2000, pp. 260-262.

[57] B. Frey, *Not Just for the Money: an Economic Theory of Personal Motivation*, Edward Elgar Publishing, Cheltenham-Brookfield 1997, trad. it. di M. Faillo, *Non solo per denaro. Le motivazioni disinteressate dell'agire economico*, Bruno Mondadori, Milano 2005.

[58] P. Pettit, *op. cit*., pp. 251-275, 302-305.

comportamento di una persona è rigidamente controllato in molti aspetti della vita (non già solo per ciò che attiene ad alcuni aspetti essenziali del vivere in comune), quando ci sono punizioni elevate, le persone, per insofferenza, per risentimento[59], per reazione, per desiderio di trasgressione, tendono spesso ad osservare le regole e le leggi non già per persuasione e *per adesione ai loro scopi*, bensì *solo perché temono le sanzioni* connesse alle trasgressioni, perciò cercano di esplorare varie modalità di violare impunemente le leggi. Così, un eccesso di controlli e di sanzioni incrementa le violazioni dei *free riders* e dunque è controproducente: accresce proprio la frequenza di quei comportamenti che vorrebbe rimuovere[60] e in tal modo aumenta anche la sfiducia in un circolo vizioso. Così, «Una volta che la sfiducia si è insinuata […] ha la capacità […] di generare una realtà [e delle condotte] coerente con se stessa»[61].

Frey sviluppa questo discorso (soprattutto nella sfera economica) valorizzando molto i risultati di quelle azioni che vengono realizzate da soggetti che agiscono sulla base di *motivazioni intrinseche*, quelle, per esempio, di chi non agisce per denaro. Un rigido sistema di controllo può rivelarsi decisamente controproducente perché sradica o comunque erode le motivazioni intrinseche e l'impegno personale per (almeno) i seguenti motivi[62]: perché riduce l'autodeterminazione e il margine di espressione inventiva e perché non di rado il controllo viene percepito come una mancanza di stima, insultante o comunque irritante.

Chi indebolisce le motivazioni intrinseche delle persone produce anomia, come nota già Mill, quando dice che le imposizioni rendono gli individui «inerti e apatici, invece che attivi»[63].

Per concludere, se Pelligra dice che «non solo l'affidabilità genera fiducia ma anche la fiducia genera affidabilità», possiamo altresì dire che l'azione virtuosa può non di rado generare fiducia e che la fiducia, tanto più se virtuosa, può non di rado generare azioni virtuose, o almeno quasi

[59] D. Gambetta, *Can we Trust Trust?*, cit., p. 220.

[60] Per esempio, un sistema fiscale oppressivo produce la moltiplicazione di strategie illegali (che, non di rado, sono in coscienza avvertite da molti come forme di legittima difesa dopo aver già pagato una certa quota di tasse) per evadere le tasse e/o aumenta il ricorso (di per sé legale) a consulenti fiscali e/o incrementa l'individuazione di attività che prevedono delle deduzioni e/o incrementa la scelta di compiere attività non tassate (per esempio viaggi e divertimenti) e/o causa il trasferimento delle imprese in altri Paesi: tutte cose che comportano risultati complessivi negativi sul gettito di un Paese, B. Frey, *op. cit*., p. 130.

[61] *Ivi*, p. 234.

[62] B. Frey, *op. cit*., pp. 19-20.

[63] J.S. Mill, *On Liberty*, Batoche Books, Kitchener 2001, p. 55.

virtuose (dipende dalle motivazioni del *trustee*). Dunque «un'apertura [fiduciale] genuina è già speranza, benché non certezza [va sottolineato], di una risposta costruttiva»[64], positiva, quasi-virtuosa o virtuosa.

6. *La fiducia più virtuosa*

Che cosa intendo dunque per atto virtuoso di fiducia? Abbiamo visto (nel § 4) che la fiducia (soprattutto nel senso I menzionato al § 2, ma talvolta anche nel senso II menzionato nel medesimo paragrafo: come detto, affidarsi all'attestazione del *trustee* può generare le risposte virtuose o quasi-virtuose del *trustee* per i motivi visti al § 4) è non di rado generativa di azione buone, abbiamo visto che la fiducia virtuosa viene concessa senza garanzie (ancorché ragionevolmente) e senza essere dovuta, dunque talvolta come dono[65] (se la motivazione non è solo o principalmente l'utilità), dunque è virtuoso *specialmente* un (ragionevole) atto fiduciale esercitato-donato anche/proprio con bene-volenza e con lo scopo di generare le altrui risposte buone; in tal senso, la virtù della fiducia si palesa come propensione a compiere-donare tali atti fiduciali in vista del bene dell'altro.

Ma, allora, visto che l'amore è (specialmente) dono e visto che vuole il bene dell'altro[66], ne segue che questi tipi di atti fiduciali sono atti dell'amore.

E l'amore fiducioso evita l'iperdirettività e il paternalismo (che è una cura malefica, anche quando è benintenzionato), non vuole conformare l'altro a propria immagine e somiglianza[67], bensì esercita il riconoscimento verso l'altro nella logica della sussidiarietà, aiutando maieuticamente la libertà altrui non già sostituendosi ad essa, bensì sostenendola, come fa appunto il sostegno con una pianta, che non la sop-pianta né la soffoca, bensì la sorregge affinché possa crescere, fiorire e fruttificare.

Del resto, la dignità della persona esige che se ne promuova l'iniziativa e la libertà[68] invece che deresponsabilizzarla e umiliarla togliendole delle mansioni.

[64] V. Pelligra, *I paradossi della fiducia*, cit., p. 263 e p. 265.

[65] Che non può essere esigito, appunto perché è un dono, e inoltre perché esigere al *trustor* il dono della fiducia significa (paradossalmente) non avere fiducia in lui, nella sua intenzionalità fiduciale benevolente.

[66] Cfr. già Aristotele, *Retorica* 2, 4.

[67] B. Mapelli, *Nuove virtù: percorsi di filosofia dell'educazione*, Guerini, Milano 2004, p. 154.

[68] In rapporto alle scelte gravemente malvagie della libertà diventano necessarie delle restrizioni; ma questo è un altro discorso.

Anzi, la fiducia non solo non poggia su garanzie vincolanti e include una misura di oggettiva incertezza riguardo all'altro, ma se vuole sapere tutto dell'altra persona si rovescia nel sospetto e rischia di rovinare, finanche sfasciare, le relazioni, come avviene tra Otello e Desdemona[69], per via della pretesa di conoscenza totale circa Desdemona da parte di Otello. Spesso la nostra letteratura ci mostra figure tragiche, il cui destino dipende dalla loro incapacità di vedere. Ma, se la cecità di Edipo risiede nel non saper riconoscere Laio e Giocasta, la cecità di Otello è diversa, perché egli «vuole troppo vedere, anche quello che non c'è» e questo gli fa perdere fiducia e «gli impedisce di vedere veramente l'amore di Desdemona»[70].

7. *La genesi della virtù ed il suo apice coinvolgono la fiducia*

Del resto, se le azioni virtuose dell'essere umano adulto sono sì state spesso generate dalla fiducia, ma non richiedono necessariamente nuova fiducia in aggiunta a quella accumulata in passato (cfr. il caso dell'eremita contemplativo virtuoso)[71] come patrimonio, viceversa la genesi delle virtù nel piccolo d'uomo richiede necessariamente la fiducia, da parte sua in se stesso e negli altri, e da parte degli altri nei suoi riguardi.

La genesi della virtù richiede fiducia verso altri, perché è vero (cfr. § 1) che la virtù ha un compito ermeneutico e inventivo, dunque il soggetto è insostituibile nella valutazione morale concreta che lo interpella e non deve delegarla ad altri. Però l'autocoltivazione morale dipende *anche* dall'educazione morale ricevuta e dalla tradizione di appartenenza: l'educazione è (anche) una trasmissione intergenerazionale di un patrimonio di esperienze e conoscenze sedimentato e non esiste un punto zero (come ha ampiamente spiegato l'ermeneutica). Ricevendo le conoscenze della propria tradizione, accettandola/respingendola in toto o in parte, il soggetto, anzitutto grazie ai genitori, apprende principi morali e virtù, li apprende in teoria e soprattutto in pratica, cioè conosce o (meglio ancora) incontra degli *exempla* di virtù (augurabilmente già i genitori), che a volte sono dei veri *phronimoi*, degli universali concreti.

[69] Cfr. B. Mapelli, *op. cit*., pp. 155-159. Qualche spunto anche in S. Cavell, *Disowning Knowledge in Seven Plays of Shakespeare*, Cambridge University Press, Cambridge 1987, trad. it. di D. Tarizzo, *Il ripudio del sapere. Lo scetticismo nelle tragedie di Shakespeare*, Einaudi, Torino 2003, pp. 149-169.

[70] B. Mapelli, *op. cit*., p. 159.

[71] Peraltro, l'eremita contemplativo è comunque nutrito da un'altra Fiducia.

Insomma, almeno all'inizio, il soggetto non può essere introdotto dall'educazione nella realtà se egli non si affida fiduciosamente (come dicono i comunitaristi ma anche un liberale come Rawls)[72] a coloro (in particolare i *phronimoi*)[73] che lo circondano e hanno cura di lui e inoltre al patrimonio di saggezza (se è tale) pratica tramandato nella sua tradizione. I piccoli d'uomo, ma non solo loro, hanno bisogno di dialogare/ascoltare fiduciosamente altri circa la vita buona e molto altro. Ovviamente un soggetto deve anche giudicare in modo critico ciò che apprende dagli altri e dalla tradizione, che non va accettata supinamente (anche la mafia è una tradizione…): il problema della dialettica delle tradizioni è molto complesso e qui non lo trattiamo.

La genesi della virtù richiede inoltre fiducia verso se stessi e da parte di altri.

Infatti, tanto per riuscire ad amarmi, quanto per esercitare una condotta virtuosa verso gli altri, ed a maggior ragione per propormi un'ideale di eccellenza virtuosa, devo nutrire autostima, devo avere fiducia in me stesso, e questa fiducia è attivata dall'altrui virtuoso e fiducioso riconoscimento. In altri termini, la mia capacità di azione virtuosa (nei miei stessi riguardi o nei riguardi di altri) è attivata, perlomeno *ab origine*, dal virtuoso e fiducioso riconoscimento altrui. Quest'ultimo è necessario per riuscire ad amare se stessi e se non riesco ad amare me stesso non sono in grado di agire virtuosamente verso gli altri: chi non si sente affettuosamente accolto non riesce ad accogliersi, non riesce ad amarsi; chi non sa amarsi non ha la carica affettiva sufficiente per esplicare le sue capacità fondamentali[74] (il tema del riconoscimento è cruciale, ma qui abbiamo potuto fare solo minimi cenni)[75]. Come dice Cruz[76], «La fiducia non serve solo come lubrificante sociale; la fiducia è segno di rispetto» ed è un'espressione di riconoscimento. «In generale, vogliamo ricevere fiducia, anche se non c'è nulla di particolare che speriamo di ottenere da questa fiducia. Infatti, essere sfiduciati

[72] J. Rawls, *A Theory of Justice*, The Belknap Press of Harvard University Press, Cambridge Mass. 1971, trad. it. di U. Santini, *Una teoria della giustizia*, Feltrinelli, Milano 1982, pp. 379-382.

[73] Ci sono *phronimoi* anonimi (saggi genitori, saggi fratelli, coniugi, amici, insegnanti, ecc.) e ci sono poi *phronimoi* eccezionali (il santo, il fondatore di un nuovo buon carisma, ecc.).

[74] F. Botturi, *La generazione del bene. Gratuità ed esperienza morale*, Vita e Pensiero, Milano 2009, pp. 163-194.

[75] Rimando a G. Samek Lodovici, *La socialità del bene. Riflessioni di etica fondamentale e politica su bene comune, diritti umani e virtù civili*, Edizioni ETS, Pisa 2017, pp. 94-105.

[76] J. D'Cruz, *Trust, Trustworthiness and the Moral Consequence of Consistency*, in «Journal of the American Philosophical Association», 1 (2015), 3, pp. 467-484, p. 482.

senza una ragione specifica e sufficiente può essere offensivo e persino umiliante. Si consideri l'indignazione e il conseguente risentimento di un cliente trattato con sospetto dal titolare di un negozio anche se non ha intenzione di rubare».

In più, l'azione virtuosa si esplica anche come giusto amore di sé (che è diverso dall'egoismo), perché io stesso merito la mia sollecitudine (si può pensare che l'amore di sé sia contrario alla virtù se lo si identifica con l'autoindulgenza, quando invece amore e indulgenza sono distinti e spesso contrari. L'amore autentico non è accondiscendente o compiacente, bensì è esigente, poiché vuole il bene, e dunque vuole anche e principalmente il bene morale, di se stessi o di qualcun altro).

Inoltre, l'esercizio delle virtù presenta spesso anche un aspetto relazionale: come dice Nussbaum[77], il vero coraggio richiede un interesse per i concittadini e per la patria; la vera moderazione richiede il giusto rispetto verso gli altri; la vera generosità richiede un interesse autentico verso il donatario, ecc.: non si possono scegliere come fini queste attività eccellenti senza scegliere come fine anche il bene di altre persone. Ora, l'apice della virtù è espressione dell'amore: il discorso sarebbe molto lungo ma, molto brevemente[78], si può qui almeno dire, accogliendo il discorso di Agostino[79] (e, in una qualche misura, modificandolo), che: la perfetta temperanza è innervata dall'amore, mi custodisce e mi preserva capace di donarmi a chi amo o, perlomeno, mi rende capace di non trattare gli altri come mezzi, bensì come fini in sé (per dirla kantianamente); la giustizia nella sua pienezza è l'amore che realizza il bene di chi amo e che gli spetta; la perfetta fortezza è innervata dall'amore con cui affronto le difficoltà, la fatica e il dolore per conseguire il bene di chi amo (me stesso o qualcun altro); la perfetta *phronesis* è sospinta dall'amore e individua in concreto le azioni che promuovono il bene di chi amo (me stesso o qualcun altro).

Più precisamente, le azioni massimamente virtuose sono quelle che esercitano l'amore (verso me stesso o verso altri) o quelle che, pur non essendo in sé azioni d'amore nella loro prima identità, sono però originate da esso e perciò acquisiscono un'identità ulteriore, diventano anch'esse atti dell'amore: come dice Agostino, «sia che tu taccia, taci per amore; sia che tu parli,

[77] M. Nussbaum, *The Fragility of Goodness. Luck and Ethics in Greek Tragedy and Philosophy*, Cambridge University Press, Cambridge 1986, trad. it. di G. Zanetti, *La fragilità del bene. Fortuna ed etica nella tragedia e nella filosofia greca*, il Mulino, Bologna 1996, p. 636.

[78] E rinviando a G. Samek Lodovici, *L'emozione del bene*, cit.

[79] Agostino, *De moribus ecclesiae contra Manicheos*, I, 15, 25, *De civitate Dei*, XV, 22.

parla per amore; sia che tu corregga, correggi per amore; sia che perdoni, perdona per amore; sia in te la radice dell'amore [indirizzato dalla retta ragione], poiché da questa radice non può procedere se non il bene»[80].

Ora, l'amore non è dif-fidente bensì è una fiduciosa promozione del bene dell'altro, nutrita dalla fiducia che l'altro possa e voglia cercare e conseguire il bene. E, come abbiamo già detto, gli atti fiduciali bene-volenti sono già essi stessi atti d'amore, i quali possono generare una risposta proprio nella forma dell'amore: «*nulla est maior provocatio ad amandum quam praevenire amando*»[81].

Ora, grazie alla loro capacità originaria di coinvolgimento, «gli *affetti* possono propriamente esercitare la loro funzione di collegamento: possono cioè configurarsi, davvero, come *legami*»[82], come legami fiduciali, creando relazioni. E, quando l'amore espresso verso l'altro ottiene una risposta cor-rispondente, tra i due soggetti si può instaurare l'amicizia, che ri-alimenta l'amore e che in modo eminente attualizza un'ontologia della relazione, secondo la quale «essere è relazione»[83].

Ebbene, la perfezione dell'esperienza intersoggettiva virtuosa, generata, almeno all'origine, dalla fiducia, è raggiunta proprio dall'amicizia, intesa come relazione di autentica e reciproca sollecitudine per il bene dell'altro. E l'amicizia, come diceva già Aristotele[84], include di nuovo la fiducia reciproca, perché gli amici si fidano l'uno dell'altro.

English title: Trust and virtue.

Abstract

The purpose of this paper is to illustrate some aspects of the relationship between virtue and trust. First, it mentions the nature of virtues and sets out a concept of reliance and trust (in its relationship with the virtue of hope). Then, it underlines the need for trust in interpersonal and social life and as

[80] Agostino, *In epistolam Johannis ad Parthos*, 7, 8.

[81] Tommaso, *Summa Theologiae*, II-II, q. 27, a. 1.

[82] A. Fabris, *Il coinvolgimento degli affetti*, in F. Botturi, C. Vigna (a cura di), *Affetti e legami*, «Annuario di Etica», Vita e Pensiero, Milano 2004, pp. 23-36, p. 33.

[83] A. Fabris, *TeorEtica. Filosofia della relazione*, Morcelliana, Brescia 2010, p. 114. Cfr. anche A. Fabris, *RelAzione. Una filosofia performativa*, Morcelliana, Brescia 2016, specialmente pp. 172-175.

[84] Aristotele, *Etica Nicomachea*, 1157a 20-25.

the foundation of the social bond, as well as its capacity (for various reasons) to generate the reliability of the trustee (even when it has already been betrayed) and, conversely, underlines the harmfulness of distrust and of pervasive controls, which tend precisely to increase those behaviors that would like to prevent.Finally, it expresses itself on the most virtuous trust, the generative one, on the relationship between trust and the genesis and the apex of virtue, which culminates in love and friendship.

Keywords: virtue; trust; phronesis; friendship; Aristotle.

Giacomo Samek Lodovici
Università Cattolica di Milano
giacomo.sameklodovici@unicatt.it

T

Trust, Implicit Attitudes, and the Malleability of Group Identities

Sarah Songhorian

1. *Introduction*

The concept of trust, just as many concepts we ordinarily use in our moral, social, and political discourse, is a complex and multifaceted one (McLeod 2015; Hawley 2012; Simpson 2012; Baier 1986). By applying it to a variety of different contexts, it is hard to have a good sense of its conceptual boundaries and of whether using it is appropriate or not in a given context. Whether, for instance, we are actually talking about the same concept when we consider self-trust, interpersonal trust, or trust in institutions is an open and difficult question to be asked. Furthermore, there is a branch of empirical research that is growing quite fast on the impact implicit attitudes and biases have on modulating trust.

The aim of this paper is underlining the necessity for a more fine-grained definition of our ordinary conception of trust that, in line with the data I will discuss, focuses on the need to distinguish clearly two different levels within the concept of trust itself. As I will show, this will also help us to get a better understanding of the role trust can and should play in ethics. Before moving to the consideration of the empirical data and to the revision of the concept of trust I will suggest, it is important to get a sense of what we ordinarily mean by trust.

Pretheoretically, trust is an important force prompting us to rely on others – human being, institutions, or authorities – and to build a relationship with them. While this general understanding of trust can be applied to all its forms and varieties, in what follows I will restrict myself to the consideration of interpersonal trust as it is the first one to be modulated by the data I will consider. This, however, does not mean that implicit attitudes

 DOI: 10.4454/teoria.v39i1.61

cannot have an impact also on other forms of trust or that this possible outcome is irrelevant, but simply that it is secondary.

Basically, A trusts B only if A relies upon B to meet her commitments – whether by doing something, by saying something, by behaving in a certain way, or by being in a certain way – and, thus, A *believes* B to be trustworthy (analogously, Hawley 2012: 6). It has to be noticed, however, that the fact that A (the trustor) believes B (the trustee) to be trustworthy does not imply that B *is* actually trustworthy. Therefore, A's trust can be unwarranted or ungrounded since it may target a non-trustworthy subject as if she were to be trusted (McLeod 2015). This feature of our colloquial understanding of trust is usually easily identified and it is because of it that trust is considered important in our interpersonal relationships but at the same time risky, since it implies that subjects rely on others for matters that are of interest to them and that those others may not deserve to be considered reliable.

This ordinary conception of trust is often taken as an obvious requirement for cooperation, and, being the latter of interest for morality, it is also taken as a requirement for morality to flourish (Baier 1986: 232). However, this does not *per se* imply that every time we trust someone we are in a moral relationship to that person. Trust can, in fact, also refer to amoral situations or interactions (as we shall see in § 2 and 3).

The data I will review in what follows (§ 2 and 3) will show another, deeper, reason to think that trust can be risky. While the possibility of trust being unwarranted or ungrounded is recognized by anyone who has a concept of trust – since the most common reason to feel one's trust betrayed is the non-correspondence between one's belief that the other is trustworthy and the other's actual trustworthiness –, the risk of implicit modulation is hardly recognized or considered. Very few people would accept that their trusting attitudes are malleable to several unconscious drives. In fact, while we often are aware of our decisions and actions taking place in a certain context, most of us are unconscious of the very *influence* such situational factors can play (Herdova 2016: 52). In order to understand exactly what aspect of trust – or what level – is modulated by these drives, I will have to provide a two-level conceptualization of trust (§ 4).

In this paper, I will restrict myself in two respects. On the one hand, I will only consider one specific unconscious drive that modulate our moral interaction with others, among the many identified in the literature – namely, group identity. Hence, I will not be interested in other well-known springs of emotional and situational modulation such as, for instance, the fact that a clean setting may decrease the severity of our moral

judgments (Huang 2014; Schnall *et al.* 2008) or that a good scent may promote reciprocity and charity (Liljenquist 2010). While these data could be used to question our colloquial understanding of trust as a stable attitude, just as much as situationism has used them to challenge the existence of stable dispositions and personality traits (e.g. Harman 1999, 2000, 2001, 2003, 2009; Doris 1998, 2002, 2010; Appiah 2008), it is not my aim in this work to follow that path. The analysis I will put forth aims at showing how the concept of trust itself requires a better conceptualization rather than at suggesting that individuals' trusting traits are unstable (which most likely are). On the other hand, I will focus only on the effect of these drives on interpersonal trust (henceforth, trust *simpliciter*). This, however, does not imply that these unconscious drives can only impact the extent to which we trust others in face-to-face interactions. Quite the contrary, they can also increase or decrease the extent to which we empathize with them (Xu *et al.* 2009), the extent to which we behave altruistically (see for instance the research on parochial altruism; Bernhard *et al.* 2006) or cooperatively (Greene 2013), and the extent to which we help others (Levine *et al.* 2005); just as much as they may have a secondary impact on other forms of trust as well. And yet, the focus on possible influences on interpersonal trust is motivated by at least two reasons. First, since trust is an attitude and not an actual behaviour, it might be that altruism and cooperation depend at least in part on our trusting attitudes and not the other way around (Baier 1986: 232 considers obvious the connection between cooperation and trust). Hence, seeing the impact of these unconscious drives on trust can reveal a more fundamental effect as opposed to focusing on behaviour. Second, trusting attitudes are easy to test and the data on them are quite clear once exposed.

2. *Group Entitativity and Social Categorizing*

As Joshua Greene points out in his *Moral Tribes*, «morality evolved to enable cooperation» (2013: 23). According to Greene, cooperation is crucial for morality. For the purposes of this work, it is important to underline how cooperation is rendered possible by trusting those with whom we cooperate. Hence, it does not seem exaggerated to claim that, in order to have cooperation, we need to have some level of trust (Baier 1986: 232). Trust is, therefore, taken to be necessary for cooperation, even though it may not be sufficient.

Greene also adds, that:

> Biologically speaking, humans were designed for cooperation, *but only with some people*. Our moral brains evolved for cooperation *within groups*, and perhaps only within the context of personal relationships. Our moral brains did not evolve for cooperation *between groups* (at least not *all* groups) (Greene 2013: 23, emphasis in original).

This insight into how our ability to cooperate is limited by our group affiliation depends on several different data. First, there is ethological evidence on animals cooperating with conspecifics but not with other species. Even more interestingly, within the same species, it is more common to see animals cooperating with kin rather than with non-kin (Hamilton 1964; Wynne-Edwards 1962; Kropotkin 1908). Second, psychological research on social categorizing has shown in the past 40 years that the behavioural influences of our recognition of a group identity are very solid. These influences can either be explicit or implicit depending on whether the subject is or is not aware of their functioning and whether she endorses, avows and self-attributes them or not (Levy 2017: 3). The more one perceives group entitativity – that is, the more one perceives a group as a real entity characterized by similarity and cohesion among its members (Plötner *et al.* 2016; Dasgupta *et al.* 1999; Hamilton *et al.* 1998; Yzerbyt *et al.* 1998) –, the more she would be likely to favour her own group members over people belonging to other groups. To detail this second line of research, I will briefly expose some of the data collected to show how cooperation and trust are implicitly modulated by social categorizing both in adults and in children. The following discussion will be grounded on the assumption, shared by Baier and Greene among others, of an obvious or at least commonly recognized relationship between cooperation and trust. While this assumption most certainly deserves a deeper consideration and calls for an argument in its favour based on independent grounds, in this context I will have to take it for granted for the sake of the argument.

When adults have to predict whether they will receive more money from an unknown allocator belonging to their same group as opposed to one belonging to another group, they strongly prefer trusting their in-group members (from 76% to 89%, see Foddy *et al.* 2009: 421), if they are told that the allocator knows about their own group membership (*common-knowledge condition*). If this last condition is not met – i.e. when participants know that the allocator is unaware of their group membership –, what matters is the stereotype associated with both in-groups and out-groups (*private-*

knowledge condition). Foddy and colleagues' (2009) subjects were economic and nursing majors. Hence, when membership was known by all participants, subjects trusted their in-groups more (or, in other words, thought in-group members to be more trustworthy); whereas when the group affiliation was clearly unknown to the allocator, subjects decided whether to trust others or not making the stereotypes associated with economic and nursing majors much more salient: «The percentage of participants who chose an ingroup allocator was larger when the out-group was economics majors (80%) than when it was nursing majors (41%)» (Foddy *et al.* 2009: 421). To control whether subjects will still prefer trusting in-groups as opposed to out-groups when they could choose a sure thing (AU$ 6.00 from the experimenter) and avoid trusting altogether, Platow and colleagues conducted two further studies (Platow *et al.* 2012). The results of these studies were in line with the data collected by Foddy and colleagues (2009): participants decided to trust in-groups even when they could have dropped out in the common-knowledge condition. Hence, the data on trusting attitudes towards in-group members were enhanced by these further researches: subjects do not only trust in-group over out-group when they had to trust someone (*relative trust*), but also when they had the opportunity to opt out (*absolute trust*).

Similar data on adults were also collected in several studies on the investment game (Stanley *et al.* 2011; Güth *et al.* 2008; Tanis, Postmes 2005; DeBruine 2002). The investment game – also known as the trust game – is an economic game often used to measure the extent to which people trust others (Johnson, Mislin 2011; Berg *et al.* 1995). In this game, subjects have to decide whether to invest the money they have earned by participating in the experiment. Once the money is invested, experimenters tell participants that the money will be given augmented – e.g. tripled in the studies conducted by Tanis and Postmes (2005) and by Güth and colleagues (2008); and quadrupled in the one by Stanley and colleagues (2011) – to another individual, who can choose whether to reciprocate or not. It is in the participants' best interest to invest as much money as possible if they trust the counterpart. In this kind of research, «the measure of trust was an ecologically relevant consequential decision about how much money to risk in each interaction» (Stanley *et al.* 2011: 7712), rather than an explicit assessment by the participants of how much they trust others or of whether they thought the counterpart was actually trustworthy. This is extremely important for at least two reasons. First, given that the presence of trust is inferred from monetary interactions, the authors can grant that the participants' actual and conscious motivation is

the desire of gaining money (i.e. it is self-interest that moves them rather than a benevolent or altruistic drive). However, in order for the subjects to gain the most money they could, they had to actually trust their counterparts. Therefore, despite participants' motivation could be taken to be mere self-interest, the authors can easily claim that the latter has to be backed with trust for the subjects to actually behave as they do: if they were not to trust the counterpart to send them an adequate amount of money back, they would decide not to invest at all precisely for their own self-interest. Second, as will become clear in § 4, the disjunction between explicit and implicit measures is best explained by the account of trust I will suggest as opposed to the current colloquial understanding of it. To elicit group membership, participants were either shown a picture of the alleged counterpart (Stanley *et al.* 2011) – either an unknown black or white individual –, or they were given information about their counterpart's personal (picture and name) and social identity (University affiliation) (Tanis, Postmes 2005). Stanley and colleagues found that «Individuals whose IAT[1] scores reflected a stronger pro-white implicit bias were likely to offer more money to white partners than black partners, and vice versa» (Stanley *et al.* 2011: 7713). Hence, this evidence corroborates the thesis according to which implicit attitudes elicited by racial cues modulate the extent of trust granted to individuals. Analogously, Tanis and Postmes conclude that «There was less trusting behaviour when the counterpart was not personally identified and a member of the outgroup» (Tanis, Postmes 2005: 419). DeBruine's (2002) version of the trust game is slightly different from the standard one, but reaches similar conclusions. In particular, DeBruine's subjects had to decide whether to divide equally a small amount of money with another participant or to trust the other participant to divide a larger sum. As in all versions of the trust game, the other participant could also choose not to divide it. Apart from this slight difference with respect to the game used, the interesting aspect of this research is that data on actual decisions were evaluated against data of facial resemblance. To create cues of kinship, pictures of participants were manipulated

[1] The Implicit Association Test (IAT) measures the strength and speed of the association between a concept and an attribute (Greenwald *et al.* 1998). Subjects see on the top of the screen two categories, one on the right and one on the left (e.g. Black and White, Male and Female), and in the middle of the screen the word of the negative or positive feature to be attributed to one of these categories (e.g. pleasant, unpleasant, good, bad, vicious, virtuous). The amount of time spent to do the association is measured and it provides an implicit measure of participants' preferences.

using digital morphing techniques, so that the pictures that were later shown to them – as pictures of their counterparts – had different degrees of similarity with themselves. DeBruine's results are in line with both the data on in-group favouritism just reviewed and with the evolutionary data mentioned earlier. Hence, not surprisingly, facial resemblance – being a cue of kinship – enhances trust and makes subjects more inclined to trust «opponents who resembled themselves significantly more than they trusted other opponents, but did not reward trusting moves by their opponents differentially» (DeBruine 2002: 1311). This last piece of evidence will be of interest in § 4 as a further element that can be more easily explained by a two-level characterization of trust, as opposed to an ordinary one.

Interestingly, children show a similar pattern of preferences towards in-group members over out-group ones from the age of five or six years:

> Many studies have shown that preschool children prefer members of their language (Kinzler, Dupoux, Spelke 2007; Kinzler, Shutts, Dejesus, Spelke 2009), gender (Martin, Fabes, Evans, Wyman 1999; Shutts, Kinzler, McKee, & Spelke, 2009), and (to some extent) racial in-groups over out-groups (Kinzler, Spelke 2011; Kinzler *et al.* 2009) (Plötner *et al.* 2015: 162).

These data show that, as happens in adults, children also display different behaviours and attitudes when they interact with people who belong to their same group, as opposed to what they do when they interact with members of other groups.

The aim of this section was to provide some insight into a research field that has been providing evidence for quite some time now on how adults and even children have a preference for trusting and cooperating with members of their own group over members of other groups. What this literature cannot tell us, however, is whether this preference depends on the fact that subjects are more familiar with their in-groups rather than with their out-groups – as claimed by Ziv and Banaji (2012) to account for children's preferences – or on something else. To solve this problem one should resort to another line of research: that of the minimal group paradigm (§ 3).

3. *The Minimal Group Paradigm*

According to the minimal group paradigm, the preference for one's own in-group holds even when the salient groups are created arbitrarily in the lab (hence, the definition of these groups as "minimal"), as well as when

subjects are given little or no time for real face-to-face interaction or when they are provided with very few cues of such a shared belonging (Plötner *et al.* 2016; Locksley *et al.* 1980; Brewer 1979; Brewer, Silver 1978; Tajfel 1974; Tajfel *et al.* 1971; Tajfel 1970). Therefore, with this line of research one can get rid of the objection according to which our preferences for in-groups may depend on familiarity. When groups are created in the lab and are not based on any visible cue, subjects are equally familiar with in-groups and with out-groups. Therefore, should the effect hold, we would need to find a reason for it without resorting to familiarity.

Plötner and colleagues (2015) have shown that 5-year-olds display a preference for members of their own minimal-group (same colour t-shirt) on multiple dimensions and even after a brief interaction. In particular, as far as trust is concerned, after children saw two puppets – one with the same colour t-shirt (in-group) and one with a different colour t-shirt (out-group) – select different boxes containing toys, they were asked to choose a box, without having the possibility to previously look into the two alternative boxes for themselves. At the age of 5, children tend to trust significantly their counterpart's choice after having cooperated with them and they show a trend to trust them more in the minimal group paradigm.

Tajfel, one of the pioneers of the minimal group paradigm, conducted several studies using this methodology (Tajfel 1974; Tajfel *et al.* 1971; Tajfel 1970). He tested 14- and 15-year-old boys with whom he pretended to divide them according to a specific criterion (i.e. he pretended to divide them among "over-estimators" and "under-estimators" of the dots that appeared on a screen, based on whether they performed "better" or "worse" at estimating the number of dots, or based on the alleged detection of a preference for Kandinsky or Klee), while they were actually divided randomly. After the division phase was completed, participants had to attribute penalties and rewards to an unknown partner – since participants were significantly older than those in Plötner and colleagues' experiment, there was no face-to-face interaction. The only thing that they were told was whether the other boy was from their own group or from the other one. Since the boys all knew each other being schoolmates, this was a tool to avoid previous friendships or hostilities to get in as confounders and to avoid any personal cue to enter the experiment. The only group identity that had to be elicited was the one Tajfel had previously given them by dividing them in two groups. As a result, participants were more kin to give more rewards and less penalties to members of their (arbitrary) group compared to the penalties and rewards they gave to members of the (arbitrary)

out-group. This evidence suggests that the implicit attitudes and the biases associated with identifying with and belonging to a certain group can be triggered easily and can be elicited also by randomly selected and arbitrary groups. And this testifies to the claim that group affiliation and social categorizing are malleable. In fact, we activate the same preferences we have for long-term and stable groups (based on ethnicity, gender and the like) also for new and rather insignificant ones (e.g. wearing the same colour t-shirt as in Plötner *et al.* 2015, or supposedly preferring Kandinsky over Klee as in Tajfel 1970).

If one is worried about the possibility that in-group favouritism and the biases associated with out-groups may lead to dehumanizing members of the latter (Varga 2017), then these data constitute at the same time good and bad news, since one could arbitrarily modify who belongs to what group. The negative aspect is clearly that in-group favouritism can and is triggered without us knowing about it, and even when there are no distinctive cues supporting it. And yet, on the positive side, this malleability of in-group favouritism can also be seen as an opportunity, rather than only as a limit. Since it is so easily triggered even by simply dividing subjects in the lab, it seems at least theoretically possible to manipulate the sense of belonging so as to include people that were previously conceived of as belonging to the out-group. The upshot is that: if one aims at enlarging the scope of people whom we trust, in-group favouritism can be used as a tool to obtain such an outcome in line with the Contact Hypothesis, according to which inter-group contact would reduce stereotyping, prejudice, and discrimination (Dovidio *et al.* 2003; Allport 1954).

4. *A Two-Level Concept of Trust*

The data reviewed in the previous sections (§ 2 and 3) show that humans naturally tend to favour in-group members over out-group ones when it comes to cooperating with and trusting them. The evolutionary reasons for this are clear: as we share the goal of preserving our kins (see on this the debate on the selfish gene; Dawkins 1976; Williams 1966), we expect in-group members to act aiming at this same goal; whereas we do not expect out-groups to share it and to act in favour of it. Hence, we believe in-group members to be more trustworthy than out-group ones. While there are evolutionary reasons for this differential attitude to be in place, one could wonder whether it is also *morally justifiable* to have it. I am not

convinced that it is the case that recognizing these preferential attitudes serves as a normative justification of their existence (cf. also Singer 2009: 61). Being aware of the existence of several implicit attitudes does not imply *per se* that we do not have a moral obligation to overcome them (de Lazari-Radek, Singer 2014).

On the contrary, recognizing that certain implicit drives can modulate our attitudes should be reason enough to focus on their influence and to try and limit it. In particular, one should wonder what the actual object of such influence is. If one takes trust as a unitary concept, then one is deemed to consider it subject to these drives in all its occurrences and forms. And that would mean that anytime we trust someone or something we are actually doing it *because of* these unconscious drives and not of appropriate reasons. Were it the case, trust would be deprived of any relevance in ethics since it would only be the manifestation of unconscious processes of which the subject is unaware. Just like a nervous tic, we would be unable to attribute moral responsibility to it. Trust would, thus, turn out to be amoral in all of its forms and occurrences. Hence, the data reviewed above would not be conceived of, as they should, as peculiar *amoral* cases of trusting attitudes being influenced by unconscious drives, but would be paradigmatic cases of what happens each and every time we trust someone even in situations that are actually morally relevant. On the contrary, recognizing that trust may have at least two different levels would improve our comprehension of the concept itself and would avoid considering it at the mercy of in-group favouritism at any time. It is for these reasons that I take a two-level conceptualization of trust to be more useful and to be able to preserve trust's moral dimension that would otherwise be lost because of the kind of data I have discussed. In what follows I will briefly describe what I take to be these two levels.

The first level is characterized by low-level, automatic, unconscious, and often even amoral trusting attitudes (like the ones reviewed above). It is at this level that social identities can play a role in modulating our responses. The second level, on the contrary, is the one that refers to conscious deliberations to trust someone. This is more cognitive, conscious, and deliberated. While the former is fast and refers to attitudes we often are unaware of – I might have no idea that the reason why I am more prone to expect reciprocity in a trust game from a certain counterpart (an in-group) rather than from another (an out-group) is that I have an implicit preference towards members of my own group –, the latter is the one we are interested in when attributing moral responsibility and when morally

evaluating the character or behaviour of an individual. If the subject was to avow and self-attribute the reason guiding her to choose an in-group member over an out-group one, then there would be grounds to morally evaluate that attitude and the actions deriving from it. That is to say that, from the standpoint of normative ethics, one could not judge automatic trust as praiseworthy or blameworthy insofar as the individual did not choose to have such an attitude and may also be unaware of it[2]. If I tend to favour women in a trust game and I am not aware of this in-group's influence, then I should not be morally judged for having such an implicit attitude. On the contrary, if one endorses and avows her own implicit attitudes, then that person can and should be evaluate morally. For instance, if, besides unintentionally behaving in a certain way towards an ethnic out-group in a trust game in the lab, I also claim that it is right to do so, and I indulge and endorse discrimination, then I am consciously and deliberately trusting some individuals more than others based on aspects that have nothing to do with someone's trustworthiness. Skin colour or gender have, in fact, nothing to do with people's trustworthiness.

This two-level account of trust is relevant for at least three reasons. First, it allows us to explain some of the empirical evidence discussed above. For instance, by claiming that one thing are our automatic and unconscious trusting attitudes and another our deliberate ones, one is capable of accounting for the fact that, in DeBruine's experiments, subjects trusted more those who physically resembled them but did not reward trusting moves differently (2002: 1311); this distinction can also account for the absence of differences in participants' explicit assessment of how much they trust others or of whether they thought the counterpart was actually trustworthy (Stanley *et al.* 2011: 7712). Both these data can be interpreted as showing that, while the effect of implicit attitudes works perfectly well and in a very direct way as far as the automatic mechanism is concerned, it does not go through when a certain amount of reasoning and deliberate behaviour is required. When subjects have to reward others or have to assess explicitly the situation, the effect of the unconscious preference for one's own in-group members decreases or disappears. Second, this account grants that normative moral theory can be concerned with the concept of trust in its deliberate form. Deliberate trust can be morally judged

[2] This clearly depends on the notion of moral responsibility at stake and on the extent of control and awareness subjects have on their implicit attitudes and on the behaviours based on them. I have dealt with this issue in another paper, see Songhorian (2018).

– if I deliberately trust only members of my own group even if I could do otherwise, then I am and should be subjected to moral judgment – and it can be cultivated in a positive and virtuous way. Deliberately trusting others is often the morally good thing to do. Interestingly enough, while deliberation has been granted a crucial role in ethical reasoning since Aristotle (*Nicomachean Ethics*, III.3.1112a-113a), little has been said concerning its connection to the concept of trust. Distinguishing between low-level and conscious trust can, thus, also help relating these two concepts to one another. Third, this account helps getting a better grasp of the notion of trust and avoiding exaggerating or underestimating the influence implicit drives can have on it, either believing that our notion of trust has to be abandoned altogether because of the effect of implicit attitudes on *some* of its occurrences or that there is no need to modify the notion itself.

This account can also serve better than the ordinary one the purpose of understanding how the minimal group paradigm can be used to increase inter-group contact and to avoid dehumanization. Social categorizing has a direct impact only on automatic and unconscious trust, as the fastness and implicit nature of the decision to be made – in a trust game for instance – stirs our ancestors' (evolutionary) reasons to unconsciously find kin, in particular, and in-groups, in general, to be more trustworthy. Unfortunately, however, social categorizing can also have an indirect effect on conscious trust. Automatically trusting more in-group members, humans tend to acquire more and more information about previous interactions with them, rather than with individuals belonging to the out-group, thus increasing the likeliness of believing the former to be more trustworthy. Discrimination can, hence, come as a conscious and deliberate endorsement of one's experience as if experience could play the role of justifying it. Rather than recognizing that the sample of people with whom one has had interactions is limited and non-representative, some may take it as good evidence in favour precisely of the option to trust those (and only those) people more. While the effect of implicit attitudes on automatic trust is necessary, their effect on deliberate and conscious trust, happily, is not and one could also realize by reasoning that there are no good reasons to prefer in-group members over out-group ones. And yet, it is at this level that groups' malleability can be of help. If subjects are unconsciously driven to experience a sense of belonging to a group – with the implicit attitudes and preferences associated to it – within arbitrary groups composed of individuals previously conceived of as out-group members, then they will have a larger sample of different people whom they have trusted automatically. From that enlarged set of previous

interactions subjects would be less likely to endorse, avow, and self-attribute explicit forms of discrimination. If one has experienced that others' trustworthiness has nothing to do with their gender or skin colour, for instance, she would less likely choose to deliberately trust only members of a certain gender or of a certain ethnic group. Group malleability can, thus, be used to directly impact automatic trust – just as much as social categorizing already does – and to indirectly impact deliberate trust – by modifying the set of previous experiences a subject has to make her inferences and to reason on in order to decide whom to trust. By showing how malleable groups are, the minimal group paradigm obtains humanization, which is the exact contrary of the dehumanization that leads to stereotyping, prejudice, and discrimination. The minimal group paradigm can be, therefore, used as a practical tool to make people from different groups enter in contact with each other (in line with the Contact Hypothesis).

5. *Conclusion*

The aim of this paper was to provide some evidence in favour of the need for a revision of our ordinary concept of trust. Evidence from studies on the investment game – also known as the trust game – and on strangers' allocation of money to an out-group or an in-group (Platow *et al.* 2012; Stanley *et al.* 2011; Foddy *et al.* 2009; Güth *et al.* 2008; Tanis, Postmes 2005; DeBruine 2002) suggest that our group identity modulates the extent to which we trust others – favouring member of our own group over members of other groups (§ 2). To get rid of the objection according to which the data collected could depend on the fact that subjects are more familiar with in-groups than with out-groups, I have resorted to the minimal group paradigm (§ 3). The evidence from this important line of research does not only eliminate the possibility of claiming that familiarity plays such a role in shaping humans' attitudes and behaviours towards in-groups and out-groups, but it also pushes the results further by claiming that the effect is not evident only in cases in which the identities at stake are strong and entrenched ones (like ethnicity or gender, for instance), but also when groups are created arbitrarily in the lab (Plötner *et al.* 2015; Locksley *et al.* 1980; Brewer 1979; Brewer, Silver 1978; Tajfel 1974; Tajfel *et al.* 1971; Tajfel 1970). Furthermore, the data from the minimal group paradigm are particularly revealing of the malleability and of the almost immediate impact social categorizing can have on automatic trusting

attitudes. While this clearly bears huge risks of malevolent manipulation, this malleability can also be an opportunity: it seems at least theoretically possible to manipulate the sense of belonging – and the automatic trust that follows from it – so as to include people that were previously conceived of as belonging to other groups. The upshot is the following: if one aims at enlarging the scope of people whom we trust to achieve humanization as opposed to dehumanization, our own biases can be used as a practical tool to obtain such an outcome.

These two lines of research have been used to show that there are several implicit drives that can modulate our trusting attitudes even if we do not know about them. Recognizing this leads to a revision of our ordinary conceptualization of trust – that I briefly discussed in § 1. Without such a revision, in fact, we would run two symmetrical risks: either exaggerating or underestimating the influence of implicit drives. If we exaggerate their impact and consider them applicable to a unitary concept of trust, then it seems that the notion of trust can have no role in ethics as there is nothing deliberate about it. On the contrary, if one aims at maintaining exactly the concept of trust we currently use – with its indistinction between amoral and moral applications –, then one needs to refute the possibility of implicit attitudes having any impact on trust altogether. To avoid these symmetrical risks, I proposed a two-level characterization of trust that would better serve the purposes of accounting for the data here discussed and for the role trust can and should play in ethics (§ 4). I have argued that the kind of trust that is subject to modulation and distortion by these unconscious drives is not the same kind we are interested in when we do moral philosophy. That is, one has to distinguish between low-level and automatic trust – the amoral one that can easily and unconsciously be biased – and more cognitive, conscious and deliberated forms of trust – those that are actually morally relevant. Since the latter is an attitude that the agent reflectively endorses and self-avows, it manifests one's moral personality and the agent can be deemed fully responsible for the actions stemming from deliberate trust. On the contrary, these features do not apply in the case of automatic trust. Hence one can conclude that, even though a moral theory would have to be primarily concerned with deliberate and conscious trust, there is still room for using our limitations to our benefit by unconsciously modulating automatic trust since it would modify the set of experiences an individual has as a basis for future inferences and expectations. This would not lead *per se* to huge differences in moral behaviour, but it might spread a positive bias that a moral agent would try to pick up and cultivate at a more conscious level.

In conclusion, in this paper I have focused in particular on the impact implicit drives have on interpersonal trust leaving aside other relevant issues that deserve more attention that I could devote to them here. For instance, I have not delved into how unconscious drives can impact other forms of trust – as trust for institutions or self-trust. Furthermore, I have not focused on whether automatic forms of trust for the dearest and nearest have *any* moral relevance or consequence *per se*. Is it a good character trait to automatically trust others? Does it lead to some interpersonal virtue? Or, on the contrary, being unconsciously trusting is completely out of the domain of moral action? While I have claimed that deliberate trust is the properly moral one, this does not imply that having a trusting character cannot have some moral consequences (as, for instance, leading more easily to some virtues like benevolence). These issues would be the object of further analysis and research.

References

Allport G.W. (1954), *The Nature of Prejudice*, Addison-Wesley, New York.

Appiah K.A. (2008), *Experiments in Ethics*, Harvard University Press, Cambridge.

Aristotle (2000), *Aristotle: Nicomachean Ethics*, R. Crisp (ed.), Cambridge University Press, Cambridge.

Baier A.C. (1986), *Trust and Antitrust*, in «Ethics», vol. 96, pp. 231-260.

Berg J., Dickhaut J., McCabe K. (1995), *Trust, Reciprocity, and Social History*, in «Games and Economic Behavior», vol. 10, pp. 122-142.

Bernhard H., Fischbacher U., Fehr E. (2006), *Parochial Altruism in Humans*, in «Nature», vol. 442, n. 7105, pp. 912-915.

Brewer M.B. (1979), *In-group Bias in the Minimal Intergroup Situation: A Cognitive-Motivational Analysis*, in «Psychological Bulletin», vol. 86, pp. 307-324.

Brewer M.B., Silver M. (1978), *Ingroup Bias as a Function of Task Characteristics*, in «European Journal of Social Psychology», vol. 8, pp. 393-400.

Dasgupta N., Banaji M.R., Abelson R.P. (1999), *Group Entitativity and Group Perception: Associations between Physical Features and Psychological Judgment*, in «Journal of Personality and Social Psychology», vol. 77, n. 5, pp. 991-1003.

Dawkins R. (1976), *The Selfish Gene*, Oxford University Press, Oxford.

de Lazari-Radek K., Singer P. (2014), *The Point of View of the Universe. Sidgwick and Contemporary Ethics*, Oxford University Press, Oxford.

DeBruine L.M. (2002), *Facial Resemblance Enhances Trust*, in «Proceeding of the Royal Society of London B», vol. 269, pp. 1307-1312.

Doris J. (1998), *Persons, Situations, and Virtue Ethics*, in «Noûs», vol. 32, pp. 504-530.

Doris J. (2002), *Lack of Character: Personality and Moral Behavior*, Cambridge University Press, Cambridge.

Doris J. (2010), *Heated Agreement: Lack of Character as Being for the Good*, in «Philosophical Studies», vol. 148, pp. 135-146.

Dovidio J.F., Gaertner S.L., Kawakami K. (2003), *Intergroup Contact: The Past, Present, and the Future*, in «Group Processes & Intergroup Relations», vol. 6, n. 1, pp. 5-21.

Foddy M., Platow M.J., Yamagishi T. (2009), *Group-Based Trust in Strangers: The Role of Stereotypes and Expectations*, in «Psychological Science», vol. 20, n. 4, pp. 419-422.

Greene J. (2013), *Moral Tribes: Emotion, Reason, and the Gap between Us and Them*, Atlantic Books, London.

Greenwald A.G., McGhee D.E., Schwartz J.L.K. (1998), *Measuring Individual Differences in Implicit Cognition: The Implicit Association Test*, in «Journal of Personality and Social Psychology», vol. 74, pp. 1464-1480.

Güth W., Levati M.V., Ploner M. (2008), *Social Identity and Trust - An Experimental Investigation*, in «Journal of Socio-Economics», vol. 37, pp. 1293-1308.

Hamilto D.L., Sherman S.J., Lickel B. (1998), *Perceiving Social Groups: The Importance of the Entitativity Continuum*, in C. Sedikides, J. Schopler, C.A. Insko (eds.), *Intergroup Cognition and Intergroup Behavior*, Erlbaum, Mahwah, pp. 47-74.

Hamilton W.D. (1964), *The Genetical Evolution of Social Behaviour II*, in «Journal of Theoretical Biology», vol. 7, pp. 17-52.

Harman G. (1999), *Moral Philosophy meets Social Psychology: Virtue Ethics and the Fundamental Attribution Error*, in «Proceedings of the Aristotelian Society», vol. 99, pp. 315-331.

Harman G. (2000), *The Nonexistence of Character Traits*, in «Proceedings of the Aristotelian Societyin «, vol. 100, pp. 223-226.

Harman G. (2001), *Virtue Ethics without Character Traits*, in A. Byrne, R. Stalnaker, and R. Wedgwood (eds), *Fact and Value*, MIT Press, Cambridge, pp. 117-127.

Harman G. (2003), *No Character or Personality*, in «Business Ethics Quarterly», vol. 13, pp. 87-94.

Harman G. (2009), *Skepticism about Character Traits*, in «The Journal of Ethics», 13, pp. 235-242.

Hawley K. (2012), *Trust: A Very Short Introduction*, Oxford University Press, Oxford.

Herdova M. (2016), *What You Don't Know Can Hurt You: Situationism, Conscious Awareness, and Control*, in «Journal of Cognition and Neuroethics», vol. 4, n. 1, pp. 45-71.

Huang J.L. (2014), *Does Cleanliness Influence Moral Judgments? Response Effort Moderates the Effect of Cleanliness Priming on Moral Judgments*, in «Frontiers in Psychology», vol. 5, n. 1276, pp. 1-8.

Johnson N.D., Mislin A.A. (2011), *Trust Games: A Meta-Analysis*, in «Journal of Economic Psychology», vol. 32, n. 5, pp. 865-889.

Kropotkin P. (1908), *Mutual Aid: A Factor of Evolution*, William Heineman, London.

Levine M., Prosser A., Evans D., Reicher S. (2005), *Identity and Emergency Intervention: How Social Group Membership and Inclusiveness of Group Boundaries Shape Helping Behavior*, in «Personality & Social Psychology Bulletin», vol. 31, n. 4, pp. 443-453.

Levy N. (2017), *Implicit Bias and Moral Responsibility: Probing the Data*, in «Philosophy and Phenomenological Research», vol. 94, pp. 3-26.

Liljenquist K., Zhong C.-B., Galinsky A.D. (2010), *The Smell of Virtue: Clean Scents Promote Reciprocity and Charity*, «Psychological Science», vol. 21, n. 3, pp. 381-383.

Locksley A., Ortiz V., Hepburn C. (1980), *Social Categorization and Discriminatory Behavior: Extinguishing the Minimal Intergroup Discrimination Effect*, in «Journal of Personality and Social Psychology», vol. 39, pp. 773-783.

McLeod C. (2015), *Trust*, in E.N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy (Fall 2015 Edition)*, http://plato.stanford.edu/archives/fall2015/entries/trust/ [September 22nd, 2018].

Platow M.J., Foddy M., Yamagishi T., Lim L., Chow A. (2012), *Two Experimental Tests of Trust in In-group Strangers: The Moderating Role of Common Knowledge of Group Membership*, in «European Journal of Social Psychology», vol. 42, pp. 30-35.

Plötner M., Over H., Carpenter M., Tomasello M. (2015), *The Effects of Collaboration and Minimal-Group Membership on Children's Prosocial Behavior, Liking, Affiliation, and Trust*, in «Journal of Experimental Child Psychology», vol. 139, pp. 161-173.

Plötner M., Over H., Carpenter M., Tomasello M. (2016), *What Is a Group? Young Children's Perceptions of Different Types of Groups and Group Entitativity*, in «PLoS ONE», vol. 11, n. 3, pp. e0152001.

Schnall S., Benton J., Harvey S. (2008), *With a Clean Conscience: Cleanliness Reduces the Severity of Moral Judgments*, in «Psychological Science», vol. 19, n. 12, pp. 1219-1222.

Simpson T.W. (2012), *What Is Trust?*, in «Pacific Philosophical Quarterly», vol. 93, pp. 550-569.

Singer P. (2009), *The Life You Can Save. How to Play Your Part in Ending World Poverty*, Picador, London.

Songhorian S. (2018), *Implicit Attitudes' Challenge to Moral Responsibility*, in «Notizie di Politeia», vol. XXXIV, n. 131, pp. 73-88.

Stanley D.A., Sokol-Hessner P., Banaji M.R., Phelps E.A. (2011), *Implicit Race Attitudes Predict Trustworthiness Judgments and Economic Trust Decisions*, in «Proceedings of the National Academy of Sciences of the United States of America», vol. 108, n. 19, pp. 7710-7715.

Tajfel H. (1970), *Experiments in Intergroup Discrimination*, in «Scientific American», vol. 223, pp. 96-102.

Tajfel H. (1974), *Social Identity and Intergroup Behaviour*, in «Social Science Information», vol. 13, pp. 65-93.

Tajfel H., Billig M.G., Bundy R.P., Flament C. (1971), *Social Categorization and Intergroup Behaviour*, in «European Journal of Social Psychology», vol. 1, pp. 149-178.

Tanis M., Postmes T. (2005), *A Social Identity Approach to Trust: Interpersonal Perception, Group Membership and Trusting Behaviour*, in «European Journal of Social Psychology», vol. 35, pp. 413-424.

Varga S. (2017), *The Case for Mind Perception*, in «Synthese», vol. 194, pp. 787-807.

Williams G.C. (1966), *Adaptation and Natural Selection: A Critique of Some Current Evolutionary Thought*, Princeton University Press, Princeton.

Wynne-Edwards V.C. (1962), *Animal Dispersion in Relation to Social Behaviour*, Oliver & Boyd, Edinburgh.

Xu X., Zuo X., Wang X., Han S. (2009), *Do You Feel My Pain? Racial Group Membership Modulates Empathic Neural Responses*, in «The Journal of Neuroscience: The Official Journal of the Society for Neuroscience», vol. 29, n. 26, pp. 8525-8529.

Yzerbyt V.Y., Rogier A., Fiske S.T. (1998), *Group Entitativity and Social Attribution: On Translating Situational Constraints into Stereotypes*, in «Personality and Social Psychology Bulletin», vol. 24, n. 10, pp. 1089-1103.

Ziv T., Banaji M.R. (2012), *Representations of Social Groups in the Early Years of Life*, in S.T. Fiske, C.N. Macrae (eds.), *The Sage Handbook of Social Cognition*, Sage, London, pp. 372-389.

Abstract

Several empirical evidences suggest that our group identity modulates our trusting attitudes, even when groups are created arbitrarily in the lab. Hence, group are malleable entities. While it clearly bears huge risks of malevolent manipulation, this malleability can also be an opportunity: it seems at least theoretically possible to manipulate the sense of belonging – and the automatic trust that follows from it – so as to include people that were previously conceived of as belonging to other groups.

I will, thus, investigate two lines of research to be used to show that there are several implicit drives that actually modulate our trusting attitudes. From this, a revision of our ordinary conceptualization of trust seems necessary. Hence, I proposed a two-level characterization of trust that would better serve the purposes of accounting for the data discussed and for the role trust can and should play in ethics.

Keywords: minimal group paradigm; trust; social categorizing; ethics.

Sarah Songhorian
Università Vita-Salute San Raffaele
songhorian.sarah@unisr.it

T

Linking Faith and Trust: Of Contracts and Covenants

Ionut Untea

1. *Introductory Discussion: Covenants of Trust and the Equal Claims of Unequal Individuals*

Trust is best illustrated in a relationship of mutual dedication to a goal that serves not only the interests but also the overall personal development of those involved. The dimension of the personal development that encompasses trust may sometimes be at least, if not more, important than obtaining specific gains. As D'Cruz observes, the need for trust emerges in a relationship even when «there is nothing in particular that we hope to gain by that trust»[1]. Moreover, «to be distrusted without specific and sufficient reason can be insulting and even demeaning», which is why the lack of trust may lead to some sort of «alienation» of individuals that ultimately warrant relationships dominated by mistrust, rather than trust[2]. If trust is needed not only as a «social lubricant» facilitating the attaining of our specific gains[3], but also as an essential ingredient of a more profound need for personal development, then a relationship of trust involves the participants' sense of equality. Indeed, being embedded in a society structured by many kinds of inequalities stemming from differences in physical, psychological, creative, or skillful performances, the participants' sense of equality in a trustworthy relationship can be taken as a legitimate claim, rather in the dimension of their equal right to personal development.

[1] J. D'Cruz, *Trust, Trustworthiness, and the Moral Consequence of Consistency*, in «Journal of the American Philosophical Association», 1 (September 2015), n. 3, pp. 467-484, p. 482.
[2] J. D'Cruz, *art. cit.*, p. 482.
[3] *Ibidem.*

 DOI: 10.4454/teoria.v39i1.62

The overlooking of this dimension of equality involved in the relationship of trust, and the unilateral emphasis of the self-interested individual, leads to the perception that trust is needed by someone rather to make use of the others' strength or skills when one cannot attain a goal by himself. The liberal view of the relationships between self-centered individuals looking primarily toward their own interests has roots in the social ontology depicted by Hobbes in his *Leviathan*. The Hobbesian approach to the trust linking those self-interested individuals is criticized by Baier, who understands it as a «male fixation» on contracts. She criticizes in Hobbes the «cool, distanced relations between more or less free and equal adult strangers»[4] and argues in favor of a more inclusive dimension of equality, one which may be extended to many categories of individuals of the social landscape: women, lovers, husbands, fathers, the ill, the very young, or the elderly[5]. Reading Hobbes, Baier indeed identifies the equality of interest operating in the Hobbesian contractual trust, but she assimilates this equality of interest with equality of status, believing that there is no place for mutual contracts between unequal participants in Hobbes's perspective. Hobbes indeed talks about a certain «similitude of passions» animating individuals, but he links the human passions with a diversity of «objects»[6]. In this way, the Hobbesian equality of interest cannot be interpreted as exclusively an equality of status, but also as an equality of aspiration to personal development, in whatever way the participants in a relationship may deem satisfactory for their passions. If this is so, the contracts between passionate individuals may acknowledge an unequal social status of the parties, and potentially an outcome that would strengthen this inequality. In spite of this, the parties still trust each other. This means that the motivations for the participants' trust do not stem from the contract itself, which does not guarantee their equality of status, but from a more fundamental need for a personal development rooted in positive or negative approaches to passions. The passionate need for developing one's personality affects equally all the individuals involved in the relationship, including the categories announced above by Baier.

Baier also disapproves of the «hypothetical Hobbesian conversions from

[4] A. Baier, *Trust and Antitrust*, in «Ethics», 96 (January 1986), n. 2, pp. 231-260, pp. 247-248.

[5] A. Baier, *art. cit*., p. 248.

[6] T. Hobbes, *Leviathan, or the Matter, Form and Power of a Commonwealth, Ecclesiastical and Civil* (ed. by R. Tuck), Cambridge University Press, Cambridge 2003, Intro., p. 2.

total distrust to limited trust»[7] and instead favors «some form» of «innate» trust. By doing this, Baier overlooks the issue that the Hobbesian «limited» trust is not necessarily limitative as to the possibilities of the manifestations of trust by individuals interacting within the social contractual framework made possible by the political power. The reasons for the «limited trust» lie in the assurance of a coherent, and predictable, behavior, and the avoidance of risky behaviors that would endanger the entire political body. As Hobbes puts it, «civil laws», which function as indicators of contractual limits, «may neverthelesse be made to hold, by the danger, though not by the difficulty of breaking them[8]. These laws facilitate by contractual ways the relationships of mutual trust that lie at the origin of the natural interactions between individuals. Hobbes makes explicit that the «mutual covenants» in the state of nature remain even more fundamental than the contracts that occur within a politically organized society. Thanks to these covenants, the laws have been made possible and «fastned at one end, to the lips of that Man, or Assembly, to whom they have given the Soveraigne Power; and at the other end to their own Ears»[9]. Therefore, the fundamental principle that holds the political body together is not the contract, which needs to be backed by some form of punitive power, but rather the covenant, which is an expression of the natural inclinations of human beings to entrust each other with a communal *telos* that surpasses any individual destiny as long as this *telos* gives meaning to one's passion for personal accomplishment.

Baier prefers seeing this natural inclination for trusting others as an «innate» disposition, and claims that it has been replaced by Hobbes with limited trust. On the contrary, Hobbes talks about a natural disposition of trusting others, which he equates with faith. Hobbes argues that «to *have faith in*, or *trust to*, or *beleeve a man*, signifie the same thing; namely, an opinion of the veracity of the man», or of «his honesty in not deceiving». In contrast, «to *beleeve what is said*, signifieth onely an opinion of the truth of the saying». Hobbes illustrates this distinction by the belief in God: whereas the belief in God brings «not onely Christians, but all manner of men» to «hold all for truth they heare him say, whether they understand it, or not», not all of them believe what it is said in «the Doctrine of the Creed»[10]. There is also a cognitive basis for the belief in the trustworthiness

[7] A. Baier, *art. cit.*, p. 242.
[8] T. Hobbes, *op. cit.*, Ch. XXI, p. 109.
[9] *Ivi*, pp. 108-109.
[10] Hobbes's emphasis. T. Hobbes, *op. cit.*, Ch. VII, p. 31.

of others but, as Hobbes argues, this cognitive aspect is highly dependent on the moral character of the believer. In his Introduction to *Leviathan*, Hobbes admits that, although he focuses on the way the sovereign «reads» in himself the passions of «mankind», any individual can read «the characters of man's heart», except that the result may vary from person to person, depending on whether «he that reads, is himself a good or evil man»[11]. This may be understood as a veiled suggestion that, as nobody, except God – the one «that searcheth hearts»[12] – has direct access to another's heart, a good person would tend to «read» another's heart in a trustworthy way, while a bad character will tend to exhibit mistrust in his fellows, being aware of his own tendencies.

Again, Baier's conception of innate trust will still apply to this relationship. But Hobbes insists on an aspect of trust as faith that gives way to trustworthy relationships that Baier's exegesis does not take into account. The article explores this suggested openness of the Hobbesian individual toward believing, or trusting, what another individual says simply on the basis of some perceived «similitude» of that individual's aspirations and one's own. The particularity of linking trust to faith helps Hobbes assert a kind of trust that goes beyond the possibilities offered by human cognitive capabilities, since a human individual, in the same way as God, can be trusted by someone, whether or not this someone understands the words or behaviors of that individual. It is this kind of trust, intimately linked with faith, that gives way to a kind of relationship that is not, or not yet, contractual, but which nevertheless implies a mutual pact. This relationship is a covenant, which is not sustained, as the contract is, by any enforcing power, but simply by the will of the parties to believe in each other and to trust each other. The «political covenant» itself[13] is one of the many possible covenants of trust in the state of nature. These natural covenants involving a faith-based trust are not defined by laws, nor can they be obstructed by them.

Without making Hobbes's exegesis the dominant dimension of my argument, but only its point of departure, I will continue with a brief reflection on the connection between the importance of the participants' faith as a basis of trust in a relationship that may involve vulnerability. This section

[11] T. Hobbes, *op. cit.*, Intro., p. 2.

[12] *Ibidem.*

[13] B.T. Trainor, *The Politics of Peace: The Role of the Political Covenant in Hobbes's «Leviathan»*, in «The Review of Politics», 47 (July 1985), n. 3, pp. 347-369.

will be followed by a more detailed discussion of the differences between contractual and covenantal relationships. In the last section I will reflect upon the way covenants of trust are an important part of the ontology of society, since covenantal trust may positively affect the behaviors and approaches of entire communities to trusting outsiders.

2. *Faith, Trust and Vulnerability*

In interpreting the relationship between unequal individuals like the «master» and the «good wife», Baier admits that the trust of the authoritarian husband can be «rational» even when he has suspicions that the wife has «strong and operative motives which conflict with the demands of trustworthiness as the truster sees them»[14]. Trust continues to remain rational «as long as the truster is confident that in the conflict of motives within the trusted the subversive motives will lose to the conformist motives»[15]. For instance, the husband has to be certain, or make sure, that «the costs to the wife» like «economic hardship» or «loss of her children» are «a sufficient deterrent»[16]. It can be observed that, at the core of the rational trust remains a feeling of «confidence» of the trustor that he is in control of the relationship. Earlier in the article Baier sees «reasonable trust» in similar terms, defining it as the trustor's «confidence» in the other's «good will», or at least in «the absence of good grounds for expecting their ill will or indifference»[17]. Thus, the kind of confidence Baier talks about, in spite of her critique of the Hobbesian perspective on contracts, remains in the framework of a relationship of trust generated between rational actors, even when these actors are not equal in power.

The character of this «confidence» remains somehow problematic, since, in order to offer «trust» to another individual, I need to have «confidence» in my own capacity for evaluating «another's possible but not expected ill will»[18]. From this perspective Baier calls trust «accepted vulnerability»[19], although it remains unclear in what way one «accepts» vulnerability itself when having «confidence» in one's own capacity for evalu-

[14] A. Baier, *art. cit.*, p. 254.
[15] *Ibidem.*
[16] *Ibidem.*
[17] *Ivi*, p. 235.
[18] *Ibidem.*
[19] *Ibidem.*

ating a person or a situation. By asserting the centrality of one's self-confidence in a relationship of trust, Baier leaves the door open to a potential replacement of the trust in another individual by a simple trust in myself, a trust in my capacities to foresee the potential vulnerabilities.

There is, however, one aspect by which Baier tries to avoid the implication of a diminished trust in another person. When talking about the relationship between the authoritarian husband and his wife, Baier admits that the husband cannot simply «rely» on her «fear». So, in order for the relationship to be considered more than mere reliance, the husband's «confidence» needs to remain open to the «hypothesis» of the wife having «some good will and some sympathy for his goals»[20]. That Baier needs to reinforce her idea of a self-confidence by the assertion of the possibility of the other's «sympathy» can also be seen in Hardin's argument that one can only express feelings of «quasi trust» in an institution or agent working for that institution if one is «confident» that the institution is reliable. For Hardin this self-confidence is a judgment based on «inductive expectations» extrapolated from «current and past actions», but this is not enough to render full trust in an institution[21]. Hardin motivates the choice of the term «quasi trust» by arguing that we cannot trust the government, the institutions, or their agents in the same way «as we might be able to trust the people we deal with on various matters»[22].

Hardin's version of self-confidence in developing trust still does not decisively indicate how trust in myself, made possible by my «confidence» in my own capacities of evaluation, can evolve toward the trust in another person. Baier's use of the husband's «hypothesis» that the wife actually cares about his own person and projects indeed seems to break the limitative self-trust. However, while its evocation seems to be a way to maintain the reasonable character of the trust, the «hypothesis» itself may be formulated on the basis of feelings and intuitions that go beyond a «rational» assessment. In this way, Baier chooses to maintain a «rational» profile of trust by the appeal to a «hypothesis» merely formulated by reason, but which may have roots in certain stances which encompass, but go beyond, rationality. Ultimately, reliance on others cannot evolve toward trust, unless someone is willing to invest some hypothetical positive appreciation of the other's inner motivations. In Hobbes's terms, if we take trust seriously as a possibility of a relationship,

[20] *Ivi*, p. 254.
[21] R. Hardin, *Trust and Trustworthiness*, Russell Sage Foundation, New York 2002, p. 156.
[22] *Ivi*, p. 158.

something that governments are not willing to do, given the huge risks, then we unavoidably have to «read» people's inner thoughts, and the outcome will depend less on what we can know or observe from past experiences with them, than on whether we are, in the depths of our heart, and beyond the moral character of our actions alone, good or bad people. In this case, trusting somebody inevitably means investing a certain amount of faith in the relationship and taking an active attitude toward accepting vulnerability.

In this case, vulnerability is not something to be expected or not, but faced, dealt with, and, if possible, defeated. Gaining one's trust may be accomplished precisely through a preliminary giving of trust and facing vulnerabilities stemming from this relationship. Potential versions of trust that are not limited by the kind of self-trust based on confidence in one's own evaluating capacities may be the «therapeutic trust», defined by Horsburgh as being based on «a belief in the possibility of stirring someone's conscience to an extent sufficient to affect his conduct»[23], or Faulkner's discussion about a belief in a friend's innocence which empowers somebody to give her or him the «benefit of the doubt» when a mountain of evidence points toward the friend's untrustworthy character[24]. In spite of their lack of emphasis on the vulnerability of the trustor, these views remain compatible with the notion that there is a certain amount of faith involved in the trustful attitude toward the other; it is thus a trust that goes beyond one's cognitive capacities for vulnerability assessment.

When talking about self-trust, I do not intend to overlook the claim that self-trust is, after all, the point of departure for trust in others[25]. I only argue that self-trust, especially when based on «confidence» on the strength of one's own cognitive abilities, may ultimately constitute a barrier to offering trust to the person in front of me. In contrast with trust, which has its starting point in self-trust, faith does not begin as faith in oneself, but starts as faith in the other, or faith in the relationship with the other, as a condition of arriving at the conclusion of a faith in one's own capacities. Abraham's status as «the father of faith»[26] did not originate in faith in his

[23] H.J.N. Horsburgh, *The Ethics of Trust*, in «The Philosophical Quarterly», 10 (October 1960), n. 41, pp. 343-354, p. 346.

[24] P. Faulkner, *Giving the Benefit of the Doubt*, in «International Journal of Philosophical Studies», 26 (2018), n. 2, pp. 139-155, p. 139.

[25] K. Lehrer, *Self-Trust: A Study of Reason, Knowledge and Autonomy*, Clarendon Press, Oxford 2002, p. 5.

[26] E. Stump, *Goodness and the Nature of Faith: Abraham, Isaac, and Ishmael*, in «Archivio di Filosofia», 76 (*Il Sacrificio* 2008), n. 1/2, pp. 137-144, p. 143.

own capacity for evaluating the trustworthy relationship between him and God, but rather in faith in God's words even if, in the manner of Hobbes's suggestion, Abraham did not fully understand all of God's plans. Nonetheless, Abraham's relationship of faith with God, at least prior to the covenants made with him, was only unilateral. Below I will explain why Abraham's covenantal relationship of faith with God becomes reciprocal, but until then I will simply maintain that, before the covenant, only Abraham's faith was operative in the relationship with God. In a relationship between two human beings, the faith infused in trust may indeed be unilateral, or it can animate both or all of the participants' trust.

The role of faith in a relationship of trust between two individuals may have many ramifications, but here I have chosen to emphasize two aspects: one would be to limit confidence in one's own capacity for evaluating whether one deserves being trusted or not; the other would be to guide the trustor beyond the concern either for individual gains, or for potential losses, and aim toward identifying and embracing a *telos* of the relationship. The decision to follow this *telos* may reveal itself as «transformative»[27], not only for the trustor, but also for the trustee, and potentially, for the wider circle of people more or less concerned by the trustful relationship. If the partners' trust for each other is only a version of self-confidence that does not break the participants' self-centered stance, then they cannot really share a *telos* that may bring them closer to each other in feelings that go beyond social quasi trust.

When interpreting Abraham's readiness to sacrifice his two sons, firstly Ishmael by leaving him in the wilderness, and secondly Isaac through his intention to burn him on the altar, because God had asked it, Stump argues that Abraham becomes the «father of faith» not because he suspends human ethics; his faith was manifested because he somehow believed that his ethics would remain compatible with God's command, and that, in spite of all the evidence, he believed in the goodness of God[28]. Although faith in God may be different from faith in a fellow human being, what remains, that is relevant for trust between individuals, from this relationship of faith in the other lies in the trustor's willingness to believe that his partner is still a good person. Thus, in the face of all the evidence, the relationship of trust will not entail a suspension of ethics, but rather its reinforcement.

[27] R. Compaijen, *Transformative Choice, Practical Reasons and Trust*, in «International Journal of Philosophical Studies», 26 (2018), n. 2, pp. 275-292, p. 275.

[28] E. Stump, *art. cit.*, p. 143.

My thinking is in the context of the readiness of the trustor to face vulnerability, and suffer from a range of risks, from minor to life-threatening, with the overall relationship still contributing to the confirmation of ethics, rather than to its suspension.

In Baier's case of the authoritarian husband, he may trust his wife neither because he relies on her fear, nor because he is confident in his own capacity to be in control of the relationship, but rather because his wife offers him her trust first, and faces all the vulnerabilities stemming from his excesses. The wife's attitude cannot be fully explained by her own rational calculations, to which there may be attached some hypothetical sympathy for her husband's goals. The good wife believes neither in her own capacity to assess potential benefits or dangers, nor in her husband's feelings of love for her. Maybe her own love for him and her sympathy for his projects or for his own person are long gone. Nevertheless, what she is still believing in is the *telos* of the relationship, the husband's commitment to this *telos*, and the positive outcome that the relationship would have for those concerned, primarily children, but also relatives, and ultimately the society's moral coherence as she sees it. It is this faith in the *telos* of the relationship with her husband that makes her trustworthy, and the faith in her husband's attachment to the same project that makes her believe in him. The wife's willingness to believe may go well beyond the husband's capacity for assessing possible gains or losses, but as long as he encounters what would appear to be her unreasonable trust, he will simply trust her in return, responding thus to the faith the wife has in his commitment to their communal project by his own faith-infused trust.

3. *Trust: Contract or Covenant?*

Abraham's relationship of faith with God seems to have been unilateral, since only Abraham needed to believe in the goodness of God, not the reverse, prior to developing a relationship of trust with the divine being. However, after Abraham becomes the «father of faith» things change: his faith triggered God's response to invest Abraham with the honor of making covenants with him: «this is my covenant with you: You will be the father of many nations» (Genesis 17:4). This response did not come as a reward for Abraham's willingness to sacrifice his sons, but actually as a response to his faith. Thus God responds to Abraham's faith by infusing divine faith in the covenant. It may sound strange to talk about divine faith, given

God's attribute of omniscience. Nevertheless, the biblical story works as an indicator that the attitude of having faith in someone does not preclude the possibility of knowledge about one's past, present, or even future behavior. As a matter of fact, God's gesture of stopping Abraham from sacrificing Isaac at the last possible moment may be interpreted as God's own sign of willingness to invest faith in Abraham, since nobody can tell for sure whether Abraham could have accomplished the commandment or thrown away the knife the moment he saw the first drops of Isaac's blood. The sign of God's willingness to believe in Abraham, his stopping of the sacrifice, is one by which God elevates Abraham's status, from a weak and untrustworthy human being, to a partner invested with the dignity of being part of a relationship with such an eminent being as God. In their covenant relationship, the inequality of the parties involved is still a reality, but the faith invested by God in his partner made Abraham a virtually equal partner in a relationship of trust. Thus, a covenant may be established between two parties that otherwise would not, or not yet, be able to have a contractual relationship. Moreover, covenants are rather based on faith and other feelings of attachment that cannot be made explicit through contracts.

Contrary to what some would expect, divine faith invested in a covenantal relationship with a human person is not so different from the human faith in another human person. As I pointed out, God's having knowledge of past, present and future behaviors of the partner in covenant is not an obstacle to having the faith in that the trustee, despite many limitations, will prove successful in coming closer to the goal of the relationship. In the same manner, the good wife knows that her authoritarian husband has certainly been unfair to her in the past and will do so again sooner or later, but this does not preclude her from infusing faith in her trustful relationship with him, since what she believes in first and foremost is not her husband's personal capacity for self-mastery, but rather his commitment to the *telos* of the project that unites their destinies and involves the destinies of others. If this is so, then the trust involved in a covenantal relationship is not the kind of trust Baier names «rational» and defines as «the absence of any reason to suspect in the trusted strong and operative motives which conflict with the demands of trustworthiness as the truster sees them»[29]. On the contrary, faith-infused trust may appear as «reasonable» even when the suspicion that the trustee will fail to meet the expectations is very

[29] A. Baier, *art. cit.*, p. 254.

strong. This is so, because the reasonable character of the faith-infused trust consists in elevating first the dignity of the trustee as a way of making her or him responsible, rather that aiming at presenting them with a task of proving their responsibility as a condition for gaining their equal dignity with the trustor.

What seems «rational» in the kind of trust presented by Baier may lead to the fact that the trustors may see themselves better morally positioned. This may be so when talking about God or good wives, but it may not be the case in other situations, even when good partners are involved. From this point of view, the requirement for the would-be-trustee to «prove» himself worthy of the trust of the other may actually manifest as a pressure on someone, sometimes unbearable to the point of psychologically discouraging him and making him more likely to err, thus to lose his capacity for self-trust. On the contrary, if one receives the trust that is rooted in faith, this is a sign that the trustor accomplishes a covenant with the trustee, a sign by which the trustor chooses to have faith in the trustee; responding to this faith by faith-infused trust in the covenant-like relationship, the trustee will find his own dignity, and thus the freedom of an equal partner, prior to accomplishing the act that is expected by the other, instead of accomplishing the act as a test for dignity. The kind of trust coming from assessing potential benefits and risks may seem more «rational» than the trust rooted in faith, but this is not an argument for the lack of reasonableness of the covenantal trust, since the «rational» trust may be in practice at best a «pragmatic» trust[30], as it may give better results than the covenantal trust. Trust as part of a covenant relationship thus invites a different kind of «reasonableness», where the trustor not only passively guides himself according to «expectations»[31], but becomes actively ready to face unexpected risks.

In spite of Hobbes's emphasis on the lack of trust between individuals within the state of nature as the potential cause of a generalized war «of every man, against every man»[32], it is also true that there is a kind of reasonableness that is evoked by Hobbes as the fundamental premise of the emergence of the political body. The popularity of the idea of a total distrust animating Hobbes's individuals in the state of nature has led to its

[30] J. Knight, *Social Norms and the Rule of Law: Fostering Trust in a Socially Diverse Society*, in K.S. Cook (ed.), *Trust in Society*, Russell Sage Foundation, New York 2001, pp. 354-373, p. 369.

[31] R. Hardin, *op. cit.*, p. 156.

[32] T. Hobbes, *op. cit.*, Ch. XIII, p. 62.

being a key source for what is called «the prisoner's dilemma»[33]. Baumgold nevertheless emphasizes that, in spite of these interpretations, trust remains at the center of Hobbes's political theory, especially with his conception of covenant. In Baumgold's view, to interpret Hobbes's theory more «accurately» is to understand that «the Hobbesian social contract originally was a covenant» by which the «incipient subjects» promised each other to «trust one another» in their collective attachment to «that body whom they had nominated as sovereign»[34]. Perhaps even more accurately, Hobbes makes clear that in the state of nature the individuals' commitment to what Baumgold and Trainor call a «political covenant»[35] is not a simple promise, but already an ongoing engagement that defines the ontology of the collective body: «this is more than Consent, or Concord; it is a real Unitie of them all, in one and the same Person, made by Covenant of every man with every man»[36].

The difficulty in interpreting the Hobbesian conception of covenant stems from the fact that there seems to be two different kinds of covenants in Hobbes's political theory: the moral covenants in the state of nature, among which the founding political covenant is only a more sophisticated version, and the civil covenants which Hobbes sees as versions of contracts. The clear difference between the two kinds of covenant lies in the fact that the covenants made after the political society is well formed are backed by political power, as are contracts[37]. This does not mean that covenants are impossible in the state of nature or in the absence of any legally binding warranty. This is the case especially because at the core of the covenantal relationship there is trust between parties. This trust, according to Hobbes, may become weakened only upon a «reasonable suspicion», but to such a degree that, in the context of a state of nature experienced by the parties as a «condition of war», the covenant is eventually considered «void»[38]. Nonetheless, as Hobbes also emphasizes, the trust between the participants in a covenantal relationship can also lead to a mutually agreed setting in which an external arbiter is endowed by the trustors

[33] A. Baier, *art. cit.*, p. 252; E. Ullmann-Margalit, *Trust out of Distrust*, in «The Journal of Philosophy», 99 (October 2002), n. 10, pp. 532-548, p. 532.

[34] D. Baumgold, *«Trust» in Hobbes's Political Thought*, in «Political Theory», 41 (2013), n. 6, pp. 838-855, p. 847.

[35] *Ivi*, p. 847; B.T. Trainor, *art. cit.*, p. 349.

[36] T. Hobbes, *op. cit.*, Ch. XVII, p. 87.

[37] *Ivi*, Ch. XV, p. 71.

[38] *Ivi*, Ch. XIV, p. 68.

with the authority of judging upon any controversy related to the application of the covenant: «Also if *a man be trusted to judge between man and man*, it is a precept of the Law of Nature, *that he deale Equally between them.* For without that, the Controversies of men cannot be determined but by Warre»[39]. Hobbes clearly states that the source of the authority of the one entrusted to judge upon the relationship of two individuals is primarily the act itself of entrusting, to which he adds the criteria according to which the judgment should be made, i.e., the «precept» of the law of nature. There is no mention here of any special power, or force, of the external judge, but simply of his quality as a trustee and, by virtue of this trust, of the judge's necessary appeal to the moral guidance of the law of nature. The authority of the law of nature, which reinforces the authority of the one entrusted to be judge, arises from the fact that, the «precepts» of the law of nature are both «written in every mans own heart»[40], and also «dictates of Reason», in spite of their not being backed, in the state of nature, by «the word of him, that by right hath command over others»[41]. In short, the appeal to the authority of the law of nature gives somebody only a moral authority, but not the legal power to enforce the application of covenants.

This means that, if the individual who has been entrusted by the participants in a covenantal relationship to evaluate the application of their covenant to their concrete problems lacks the backing of his judgment by a reinforcing power, his judgment may easily be disregarded and become «void» in the same way as the covenant that the entrusted judge is supposed to save. Why then this gesture of the parties to appeal to a judge who has only authority, but not reinforcing power? Hobbes suggests that the entrusted judge's appeal to the law of nature, defined further on as «equity», may solve the «Controversies of men» which otherwise can only be solved by war[42]. Moreover, it can also be understood that the trustee's position of outsider in regard to the covenantal relationship of the trustors makes this individual more likely to appeal to the law of nature in an unbiased way, that is, without his reasoning being affected by self-interest in the application of the moral laws, or «precepts» of nature. On one hand, the act of entrusting an external judge, even with an authority that can easily be rendered void, may increase the likeliness of people to keep the

[39] Hobbes's emphasis. *Ivi*, Ch. XV, p. 77.
[40] *Ivi*, Ch. XLII, p. 282.
[41] *Ivi*, Ch. XV, p. 80.
[42] *Ivi*, Ch. XV, p. 77.

covenants made, as it will further the ideal of trustful relationships guided by the authority of the laws of nature which are binding both in heart and conscience. On the other hand, the event itself of local entrusting of external individuals with a special authority functions as a preliminary covenantal effort that would eventually give way to the founding political covenant, where the judge would have an authority not only based on trust and the laws of nature, but also backed by a reinforcing power.

This special pedagogical role of trust for individuals to navigate between self-interests and vulnerabilities in a state of nature in such a way as to «come together»[43] and avoid collisions of their competing ambitions[44], does not end with the formation of the political body. Even in the case where the political power is «sufficient to compell» individuals to keep their covenants and contracts[45], it does not mean that there is no place for a natural trust, i.e., more fundamental than a social quasi trust, between the parties of a covenant or contract made under the authority guaranteed by the political power of the sovereign. Even when Hobbes talks about the second kind of covenant, which has been integrated in contractual relationships after the founding of the political community, he describes it as a legal relationship where «one of the Contractors, may deliver the Thing contracted for on his part, and leave the other to perform his part at some determinate time after, and in the mean time be trusted»[46]; This means that, in spite of their being integrated in the contractual law of a society, covenants still conserve a natural relationship that cannot be completely reconstructed by the insertion of the ordering political force. This is the relationship of trust, which is richer than a mere relationship of bounding duties between the individuals compelled by the power of the Leviathan.

Hobbes makes clear that at the core of this relationship of trust, whether covenants or contracts are involved, there is an unavoidable dimension of interpersonal faith: «he that is to performe in time to come, being trusted, his performance is called *Keeping of Promise*, or Faith; and the fayling of performance (if it be voluntary) *Violation of Faith*»[47]. The two essential dimensions of the covenant, entrusting somebody to perform an

[43] *Ivi*, Ch. XIX, p. 94.
[44] *Ivi*, Ch. II, p. 5; Ch. XI, p. 48; Ch. XIV, p. 68.
[45] *Ivi*, Ch. XIV, p. 68.
[46] *Ivi*, Ch. XIV, p. 66.
[47] Hobbes's emphasis. *Ivi*, Ch. XIV, p. 66.

action in the future and the faith in that person's capacities to perform the desired action remain present in both versions of the Hobbesian covenant, the founding political covenant and the later civil covenants. In the case of the founding political covenant, it is faith, not the submission to a political power, that pulls the individuals out of the «solitary, poore, nasty, brutish, and short»[48] existence in the state of nature, since the political power had to be generated first by covenantal agreement, which required a minimum of trust for the individuals to «come together»[49]. By manifesting faith toward each other, the individuals in the state of nature elevate their own statuses: they are simple humans deemed to fail and are in their passions similar to animals like «Lyons, Bears, and Wolves»[50], but at the same time they aspire in their passions to elevate their status to something closer to that of divine beings as participants to the purity of the moral law. As Hobbes puts it, «the *Pacts* and *Covenants*, by which the parts of this Body Politique were at first made, set together, and united, resemble that *Fiat*, or the *Let us make man*, pronounced by God in the Creation»[51].

This last suggestion may be interpreted as an echo of the formula «*Man to Man is a kind of God*», which Hobbes had described in the English version of *De Cive* as expressing «some analogie of similitude with the Deity, to wit, Justice and Charity, the twin-sisters of peace»[52]. Even though Hobbes's political theory gives far more attention to the contrasting formula «*Man to Man is an arrant Wolfe*»[53], this happens, as he further explains, because «Good men must defend themselves by taking to them for a Sanctuary the two daughters of War, Deceipt and Violence»[54]. Hobbes argues that it is the individuals' goodness of heart and connection of reason to the moral character of the laws of nature that makes them, at least by heart and conscience, similar to the «Deity» and predisposed to open themselves to others with a kind of faith-infused trust, even when there is no guarantee that the trustees will perform according to the engagements made. As Ryan points out, for Hobbes «breach of covenant is like what

[48] *Ivi*, Ch. XIII, p. 62.
[49] *Ivi*, Ch. XIX, p. 94.
[50] *Ivi*, Ch. IV, p. 12.
[51] Hobbes's emphasis. *Ivi*, Intro., p. 1.
[52] Hobbes's emphasis. T. Hobbes, *De Cive. The English Version: Philosophical Rudiments Concerning Government and Society* (ed. by H. Warrender), Oxford University Press, Oxford 1987, p. 24.
[53] Hobbes's emphasis. T. Hobbes, *op. cit.*, p. 24.
[54] T. Hobbes, *op. cit.*, p. 24.

logicians call absurdity»[55]. Nevertheless, it seems that, under the perception of a «reasonable suspicion», «good men» are indeed pushed toward unreasonable fear of others, a situation which calls for the establishment of a covenant that would make possible the communal endowment of a kind of a public trustee that would be invested not only with moral authority, but also with political power. Unreasonable fear of death makes the political covenant pre-eminent, in the state of nature, over all other local covenants based on faith-infused trust.

But how may this preeminence of the political covenant occur, since covenants based on trust, in contexts dominated by fear of death, are void? As Ryan observes, Hobbes is aware of the problem that he himself generated: «It seems that to establish a power that can make us all keep our covenants, we must covenant to set it up, but that the covenant to do so is impossible to make in the absence of the power it is supposed to establish»[56]. It seems that there is an apparent gap between the covenant based exclusively on trust in the state of nature, and a covenant based rather on fear of punishment in the newly organized political body. Still, for Hobbes these two covenants are one and the same, as if the first covenant is transformed into the second by the choice of the participants. For the gap between the two dimensions of the political covenant to be filled, and for this transformation of a covenant based exclusively on mutual trust into a covenant warranted by an external power to become actually possible, it is necessary for a leap of faith to occur. There are two levels on which faith operates in such a way as to render the original political covenant possible and durable: first, at the level of the choice of the individuals to trust each other against all suspicions, and second, at the level of their faith in a future savior. At the first level, personal faith as part of a trustful relationship between the individuals in the state of nature leads to a «transformative choice»[57] in the sense that the individuals reciprocally elevate their statuses by considering themselves already virtually equal citizens, although the actual elevation of their status from mere «wolves» to equal citizens will come later, after the imposition of the political power backing the contracts. Faith in their virtual capacities makes them able to see themselves

[55] A. Ryan, *Hobbes's Political Philosophy*, in T. Sorrel (ed.), *The Cambridge Companion to Hobbes*, Cambridge University Press, Cambridge 1996, pp. 208-245, p. 225. See also T. Hobbes, *De Cive*, cit., p. 63.

[56] A. Ryan, *op. cit.*, p. 226.

[57] R. Compaijen, *art. cit.*, p. 275.

not as they are, but as they will, or should, be. Therefore, the individuals in the state of nature come to trust each other in the performance of the political covenant, in spite of a great deal of evidence that instead points toward the high probability of endangering their own lives in their chasing of an improbable ideal of humankind coherently working together. At the second level, personal faiths put together contribute to a collective expectation, defined by Hobbes as «*Salus Populi*»[58], a term with religious connotations, approaching the political project to that of an earthly salvation.

It is this «people's safety» that forms the main «Businesse»[59], or «end» legitimating the «Office of the Soveraign»[60]. While the office of judge of the one who, in the state of nature, is entrusted by the participants in a covenant is sustained by the appeal to the moral authority of the law of nature, the office of the sovereign remains anchored in this kind of authority, since «the safety of the People, requireth further, from him, or them that have the Soveraign Power, that Justice be equally administered to all degrees of People»[61], but at the same time the administration of justice is rendered more expedient by its imposition through political power. Nevertheless, despite its power, and in spite of the fact that for Hobbes the covenant is not made directly with the sovereign[62], the individual(s) exercising the office of the sovereign are still morally bound by the *telos* of the whole political body, which is *salus populi*. Both Baumgold and Trainor emphasize that the sovereign's future activity is prescribed by the covenant in spite of the sovereign's not being called to display any sign of engagement as part of the covenant between individuals. Baumgold emphasizes that the individuals, when covenanting with each other, already display their «trust» that the sovereign (individual or assembly) «will do its part, as they will do theirs»[63]. Trainor argues that, in Hobbes's perspective, the sovereign may be disobeyed by his subjects when it «acts in such a way as to directly frustrate the end of the covenant»[64]. As an expression of the combined faiths of the individuals that are still in the state of nature, the *telos* of the political covenant binds them together to the extent of becomig a «real unity», a political body invested with full life and autonomy.

58 T. Hobbes, *Leviathan*, cit., Intro., p. 1.
59 *Ivi*, Intro., p. 1.
60 *Ivi*, Ch. XXX, p. 175.
61 *Ivi*, Ch. XXX, p. 180.
62 *Ivi*, Ch. XVIII, p. 89.
63 D. Baumgold, *art. cit.*, p. 847.
64 B.T. Trainor, *art. cit.*, p. 353.

As an expression of the moral authority of the laws of nature, the *telos* of *salus populi* is set up in the state of nature, but continues to be binding «*in foro interno*»[65] upon the sovereign and upon everyone exercising any legally-guaranteed official, or social, role. Its ramifications, and the ramifications of the laws of nature within the contractual structures offered by the state remain morally binding, since Hobbes maintains that any citizen is still bound *in foro interno* by the laws of nature, a moral dimension which requires that the laws be followed not only according to the actions prescribed by laws, but rather according to their moral «purpose»[66].

The argument of this section has emphasized that Hobbes's focus on faith-related trust as part of the original political covenant in the state of nature may not be as radically different from Baier's idea of innate trust as Baier would think. This aspect becomes even clearer when bringing into focus Hobbes's opinion that the laws of nature «are not properly Lawes», but rather «qualities that dispose men to peace, and to obedience»[67]. Baier also identifies a similar kind of predisposition, or «tendency» in children to «initially impute goodwill to the powerful persons on whom they depend»[68]. The individuals in the state of nature may not be like infants, and the extent to which their faith-based trust is innate remains debatable. Nevertheless, on the first level of the expression of mutual faith-based trust, the participants in the original political covenant indeed tend to attribute goodwill to their partners in covenant when they perceive their concerted contribution as equivalent to God's *fiat*. Moreover, at the further level of the development of their concerted faiths into a collective conscience of one single political body, the individuals imagine that the all-powerful mortal god, the Leviathan, will offer an earthly salvation by displaying goodwill toward them as citizens, and lack of mercy for those that will still place themselves in the state of nature after the artificial body becomes operational.

4. *Concluding Reflection: Covenantal Trust and Social Ontology*

In Hobbes's political theory, his choice of placing the emphasis on the kind of covenants which are backed, like the contracts, by political authority, rests in the fact that the sovereign, with the notable exception of a

[65] T. Hobbes, *op. cit.*, Ch. XV, p. 79.
[66] *Ibidem.*
[67] *Ivi*, Ch. XXVI, p. 138.
[68] A. Baier, *art. cit.*, p. 242.

conqueror, cannot afford to face the potentially devastating risks entailed by the faith-based trust that characterize natural covenants. Baumgold focuses on sovereignty «by acquisition» in order to argue for the possibility of trust in Hobbes's theory[69]. The image of an all-mighty conqueror willing to place his faith in subjects rather than keeping them in a condition of slavery indeed makes sense in the framework discussed above: like the almighty and all-knowledgeable God of Abraham, the high risks of placing faith in those that were just conquered by mere force are very well known, but still the sovereign prefers to infuse a little faith in them by elevating their status, from mere slaves to subjects, as a premise of their virtual behavior as politically trustworthy members of the newly formed political body[70]. My argument has not focused on sovereignty «by acquisition», but rather on sovereignty «by institution»[71] in order to enlarge the framework of identifying possibilities of trust in the Hobbesian political theory. This wider framework has allowed me to emphasize that covenantal trust indeed makes the trustor vulnerable to potential risks; that is why the trustor needs to face these risks by placing faith in the trustee, or in the *telos* of the relationship with the trustee. Otherwise, only a concern for eliminating potential vulnerabilities will transform the covenantal trust in a contractual relationship.

Although Hobbes implies that the conqueror's faith in the virtual capacity for the conquered to behave as subjects seems more reasonable than a continuous effort to secure their subjection by fear, it appears that in the case of the sovereignty by institution the risk of vulnerability of the political body seems higher. This may be so because the faith-based trust invested by the conqueror is supposed to be only a temporary solution that would rekindle the people's trust in the new political authority, but once this happens, it is more likely that the contractual trust will take over the covenantal one, as the contractual trust seems better fitted to contribute to a framework of universally applied rules of behavior that would sustain the durability of the social relationships. In this way, being confronted with the Hobbesian issue of the people's oscillating between «too much trust» and «too much diffidence»[72] the political authority summons the subjects' social trust by contractual rules of behavior that are backed by political power. Thanks to the reinforcing power, relationships of trust remain open,

[69] D. Baumgold, *art. cit.*, p. 839.
[70] T. Hobbes, *op. cit.*, Ch. XX, p. 104.
[71] *Ibidem.*
[72] *Ivi*, Intro, p. 2.

as in the state of nature, to the possibility of being broken easily, but this time there is an added danger associated with them[73].

Liberal political ontology influenced by the image of the political body as a «unity» of all those that live under the power and authority of a political power has inherited the Hobbesian hesitation in prescribing a faith-based trust as the point of intersection of human relationships. The contract backed by the authority of the law reinforced by political power seems more reasonable than the faith-based trust which opens the way for the state's vulnerability. Even in an age when «sovereignty by acquisition» is much less frequent than in Hobbes's times, the Hobbesian solution of talking about the covenantal trust rather as a preliminary step toward achieving a contractual trust may be identified in the assumption meant to appease, in the process of elections for instance, the discontent of those who backed a candidate who did not secure enough votes. Indeed, individuals who, having agreed to play the democratic game, then lose – maybe by a very narrow margin – have to follow a different political will backed by a relatively larger number of fellow citizens. They are not slaves, but they still need to accept a covenant of faith that the newly elected leader will behave as the leader of the entire people, not only of those who elected them, while the leader needs to make this explicit by sufficient signs, like declarations, discourses, solemn ceremonies of investiture, and concrete policies.

The fact that in the contemporary world there are several cases of political instability with potential political upheaval that can result in secessionist tendencies shows the importance of reasserting political covenants. These covenants may work in several directions, for instance between citizens, between them and the political leaders, and between leaders and the citizens frustrated by the leaders' behavior in key moments of the life of the political body: these include elections, and contestations of authoritarian regimes. Covenantal trust may also be operating in the case of public critiques of political corruption, efforts to counter populist tendencies of governments, critiques of the failures of integrating refugees or immigrants or in debates surrounding the national administrations' failure to stop the process of 'brain drain', i.e. temper the exodus of skilled workers and intellectuals. Faith-based covenants of trust may be used, and some are already operating, to counter discourses that ostracize some inhabitants of a region or a country by emphasizing the potential vulnerabilities of the entire country. Covenantal trust strengthens the sense of trust of the citizens

[73] T. Hobbes, *Leviathan*, *cit.*, Ch. XXI, p. 109.

in their political leaders and administrations, and facilitates the path toward communal finding of policies that could be considered reasonable by all citizens and residents to face potential vulnerabilities and risks with long-term impact on the entire political body.

The low level of encouraging the formation of ingredients leading to covenants of faith-based trust in a society may be translated into a social bias for the multiplication of contracts. As Bacharach and Gambetta argue, multiplication of signs may make the task of «mimics» harder, since one may fake behaviors according to a limited number of indicators, but it cannot generate a behavior that successfully integrates all the signs of a particular genuine behavior[74]. This may be the case, but the multiplication of signs may also be translated into a social inclination to deal with a quality-based lack of trust in a quantitative way. The commitment of a majority to rely upon an ever increasing number of signs in order to verify the trustworthiness of some individuals that deliberately place themselves outside the community, e.g., as mimics or free riders, may have a negative impact on the perception of those that have been outsiders for reasons outside their control, e.g., immigrants, refugees, handicapped persons, women in some social contexts and many types of minorities. That is why the low level of a community's capacity to uphold covenantal relationships may be seen in the high level of social pressure surrounding the upholding of a series of signs that define the identity of that community. Hence the tendency of majorities to treat outsiders that do not manifest their conformity with the social manners, language, or customs, as potential threats. The imagined failures of perceived outsiders may be translated into pejorative labeling, as throughout history many have been deemed as sinners, fools, heretics, witches, enemies of the faith, or enemies of the people, to name but a few.

Covenantal trust infuses faith in members of those minorities that are usually suspected of breaking contractual trusts, and declares their dignity as equal partners in the collective covenant of personal development without having to first prove themselves worthy of the majority's trust. It is rather a lack of a community's capacity to face collectively the risks stemming from placing «too much trust» in someone, which makes that community revert to «too much diffidence»[75] and appeal to the quantitative signs meant to elicit contractual trust. The result of this defensive attitude

[74] M. Bacharach - D. Gambetta, *Trust in Signs*, in K.S. Cook (ed.), *Trust in Society*, Russell Sage Foundation, New York 2001, pp. 148-184, p. 172.

[75] T. Hobbes, *op. cit*., Intro, p. 2.

can be seen in the estrangement of many minority communities or individuals from the *telos* of a majority and, as D'Cruz emphasizes, even increase the likeliness of adopting the expected untrustworthy behaviors[76]. If this is so, current societies need to widen their «moral circle» of trust[77] by becoming more aware of the trust rooted in faith. Such faith still maintains a reasonable character, in spite of the perceived practicality of trust based on contracts as «rational»[78]. In this respect, if the contemporary predilection for the appeal to reason in public debates has marginalized the faith in God, it does not mean that a *reasonable* public discourse about trust necessarily entails the loss of faith in each other.

Abstract

Trust is so intimately linked with faith that sometimes trust needs faith to unfold in a relationship. I argue that the role of this faith element in trust is to elevate the status of the one in which we trust so as to emphasize the equal dignity of all the participants in the relationship of trust. Against views that focus on a «rational» trust based on an exaggerated emphasis on the capacity of self-trust as a point of departure for the trust in others, the essay develops toward the depiction of a kind of trust that is rooted in faith and still maintains a «reasonable» character. By way of discussing the implications of Thomas Hobbes's reflections on covenants and contracts, and Annette Baier's critique of what she sees as the Hobbesian «fixation» on contracts, I argue toward the identification of what I call a «covenantal trust» in contemporary political ontology.

Keywords: trust; faith; vulnerability; covenant; Hobbes; Abraham; state of nature; political ontology.

Ionut Untea
Southeast University
untea_ionut@126.com
ionutz1tea@yahoo.com

[76] J. D'Cruz, *art. cit.*, p. 482.

[77] G.I. Hofstede, *The Moral Circle in Intercultural Competence: Trust Across Cultures*, in D.K. Deardorff (ed.), *The SAGE Handbook of Intercultural Competence*, SAGE, London 2009, pp. 85-99, p. 85.

[78] A. Baier, *art. cit.*, p. 254.

II.
Philosophy, Knowledge, and the Sciences

II.
Filosofia, conoscenza e riflessione scientifica

a cura di
Giovanni Scarafile

T

Law and its Imitations in Plato's *Statesman*

Paolo Crivelli

The concept of imitation plays a central role in many areas of Plato's philosophy. It does so, in particular, in the *Statesman*'s reflections on issues of philosophy of politics. In this dialogue, Plato identifies the art of the statesman, or statesmanship, with a highly specialized branch of knowledge. To this knowledge he attributes the highest authority in the state. He goes as far as to claim that only a regime based on statesmanship is a genuine constitution. To describe the relationship of present-day regimes, i.e. the regimes that have been realized until now, to the regime based on statesmanship, Plato resorts to the concept of imitation: present-day regimes imitate the regime based on statesmanship. However, he applies the concept of imitation not only to the relationship of present-day regimes to the regime based on statesmanship, but also to that of present-day politicians to the genuine statesman and to that of law to statesmanship: present-day politicians imitate the genuine statesman and laws imitate statesmanship.

These three applications of the concept of imitation are reciprocally connected. Plato explicitly argues that present-day politicians imitate the genuine statesman *because* present-day regimes (which are ruled by present-day politicians) imitate the genuine constitution (which is ruled by the genuine statesman). In this study I explore the possibility of crediting Plato with the further claim that present-day regimes imitate the genuine constitution *because* laws (on which present-day regimes rely) imitate statesmanship (on which the genuine constitution relies).

TEORIA 2019/1 DOI: 10.4454/teoria.v39i1.63

1. *Present-day regimes as imitations*

In the *Statesman*, at 292A5-D1, the dialogue's main speakers, i.e. the Visitor and Young Socrates, agree that the only criterion for deciding whether a regime is 'correct', i.e. whether it is a genuine constitution, is its reliance on the specific form of knowledge which is statesmanship, the statesman's knowledge. Thus, the only genuine constitution is the regime that relies on statesmanship. The Visitor adds:

T1 ΞΕ. ὅσας δ' ἄλλας λέγομεν, οὐ γνησίας οὐδ' ὄντως οὔσας λεκτέον, ἀλλὰ μεμιμημένας ταύτην, ἃς μὲν ὡς εὐνόμους λέγομεν, ἐπὶ τὰ καλλίω, τὰς δὲ ἄλλας ἐπὶ τὰ αἰσχίονα [μεμιμῆσθαι][1]. 293E3 E5 293E6

VIS. As for all the others that we say ‹are constitutions›, one must say that they ‹are› not genuine nor really being ‹constitutions›[2], but imitating this one [*sc.* the regime based on statesmanship], those we speak of as well-governed[3] for the better, the others for the worse (Pl. *Plt.* 293E3-6).

The Visitor returns later to the claims of T1 by saying that

T2 ΞΕ. ... τὰς δ' ἄλλας μιμήματα θετέον, ὥσπερ καὶ ὀλίγον πρότερον ἐρρήθη, τὰς μὲν ἐπὶ τὰ καλλίονα, τὰς δ' ἐπὶ τὰ αἰσχίω μιμουμένας ταύτην. 297C1 297C4

VIS. … the others we must put down as imitations, as was said a little earlier, some of them imitating this one for the better, others for the worse (Pl. *Plt.* 297C1-4).

The only variation between the two passages worth highlighting is that in T2 the noun 'imitation' ('μίμημα', 297C2) occurs while T1 contains a form of the verb 'to imitate' ('μιμεῖσθαι', 293E4) at the corresponding point (the imitations, not their authors, are described as 'imitating').

[1] I adopt Stallbaum's expunction of 'μεμιμῆσθαι', accepted also by Burnet, which avoids a change of construction in the middle of the sentence: cf. Stallbaum 1841: 268-269; Burnet 1900-07, *ad loc.*

[2] On the basis of the 'πολιτείαν εἶναι ῥητέον' in the immediately preceding line (293E2), I supply two occurrences of 'πολιτείας εἶναι' after the two occurrences of forms of 'λέγειν' at E3 and E4.

[3] The expressions 'εὔνομος' and 'εὐνομία' are employed both in cases where good laws are present and in cases where the quality of being law-abiding is displayed: cf. Ast 1835-38, *s.v.* 'εὐνομία' and 'εὔνομος' (I 853-54) (for the second use, cf. *R.* 4. 425A3; *Sph.* 216B3; *Lg.* 2. 656C5). The English 'well-governed' has a similar extension.

2. *What sort of 'imitations' are present-day regimes?*

The Visitor does not justify his rather counter-intuitive claim that present-day regimes 'imitate' the genuine constitution. There are two possible reconstructions of his grounds for making this claim.

According to the first reconstruction, the Visitor relies on an 'ontological' use of 'to imitate' ('μιμεῖσθαι'), a use that does not involve any sort of intentionality: the idea he intends to convey by using a form of 'to imitate φ', the Visitor could also convey by using the corresponding form of 'to be a downgraded form of φ' or 'to be a surrogate of φ' (where 'φ' is a schematic letter that may be replaced with any grammatically suitable expression)[4]. According to the first reconstruction, the Visitor's reason for claiming that present-day regimes imitate the genuine constitution is the following: the only genuine constitution is the government based on statesmanship; present-day regimes are not genuine constitutions because they are not based on statesmanship; therefore, they are downgraded forms of the genuine constitution; hence, they imitate the genuine constitution. This first reconstruction has a weak spot: the argument it attributes to the Visitor is invalid because the claim that present-day regimes are downgraded forms of the genuine constitution does not follow logically from the claim that they are not genuine constitutions (many things are not genuine constitutions without being downgraded forms of the genuine constitution).

According to the second reconstruction of the Visitor's grounds for claiming that present-day regimes 'imitate' the genuine constitution, the Visitor relies on an 'intentionally loaded' use of 'to imitate' ('μιμεῖσθαι'): the idea he intends to convey by using a form of 'to imitate φ', the Visitor could also convey by using the corresponding form of 'to appear to be φ without being φ' or 'to instil the illusion of being φ' (this use of the verb is intentionally loaded because of the intentionality involved in the concept of appearance)[5]. According to the second reconstruction, the Visitor's reason

[4] Cf. Hirsch 1995- 185; Pradeau 2009: 114. An ontological use of expressions linked to 'μιμεῖσθαι' is perhaps attested in the *Statesman*'s myth: cf. 273E12; 274A2; 274D7. In some cases, imitations not only do not involve viewers, but they even actually are what they imitate (cf. Marušič 2011: 222-223). For instance, in Euripides' *Electra* Clytemnestra justifies her betrayal of Agamemnon by saying that 'whenever a husband goes astray by rejecting his marriage-bed at home, the woman is likely to imitate [μιμεῖσθαι] her husband and acquire another lover' (E. *El.* 1036-38): the imitation Clytemnestra is speaking about is not aimed at a viewer and actually is what it imitates (the wife imitates her husband who is betraying her by actually committing a betrayal).

[5] For the connection between imitating and appearing, cf. *R.* 10. 601A4-B2; *Sph.* 267A6-8. In the *Sophist* (at 234C5-6) the Visitor says that the 'images [εἴδωλα]' (234C5) of true statements

for claiming that present-day regimes imitate the genuine constitution is the following: the only genuine constitution is the regime based on statesmanship; present-day regimes are not constitutions because they are not based on statesmanship; however, they appear to be constitutions (this is shown, among other things, by the fact that 'we say' that they 'are constitutions', 293E3 – later I shall examine a further justification of the claim that present-day regimes appear to be constitutions); therefore, present-day regimes imitate genuine constitutions[6]. An analogous argument could be developed with reference to suitably shaped and polished pieces of glass: the only genuine diamonds are the bodies that have such-and-such a chemical structure; suitably shaped and polished pieces of glass are not diamonds because they do not have such-and-such a chemical structure; but they appear to be diamonds (this is the reason why they are often worn); therefore, they imitate genuine diamonds. This second reconstruction also has a weak spot: one of the premisses of the argument it brings up (specifically, the premiss to the effect that present-day regimes appear to be genuine constitutions) does not occur explicitly in passages T1 and T2.

It is difficult to choose between these two reconstructions of the Visitor's grounds for claiming that present-day regimes imitate the genuine constitution. The main difference between them is that the second reconstruction attributes a role to the concept of appearance whereas the first ignores it. Now, the concept of appearance is operative shortly before T1 (the participle 'δοκοῦντας', 'seeming', is applied to rulers at 293C8, only 12 lines before T1). Moreover, the concept of appearance is relevant to the broad context of passages T1 and T2. For, these passages are bits of an extended argument whose aim is to establish that the present-day politician is 'the greatest beguiler of all the sophists and the most expert in their art' (291C3-4)[7]. But, the sophist's art is the art of appearing to have knowledge without having it. Thus, a reconstruction of the Visitor's position that attributes a role to the concept of appearance is more plausible than one that

produced by the sophist lead certain inexperienced youths to 'judge [δοκεῖν] that truths are being stated' (234C6). The verb 'δοκέω', which here means 'to judge', can also mean 'to seem' (cf. LSJ *s.v.* 'δοκέω' I and II). This suggests that the inexperienced youths *judge* that truths are being stated in that it *seems* to them that truths are being stated. In the *Sophist*, the verbs 'φαίνεσθαι' and 'δοκεῖν' are used as equivalent variants (cf. below, n. 51 and text thereto), and there is no reason to doubt that their equivalence holds also in the *Statesman*.

[6] I am not making the (false) claim that 'to imitate φ' has 'to appear to be φ without being φ' as one of its lexical meanings. I am making the weaker claim that 'to imitate φ' can be used to convey the idea that could be more properly expressed by using 'to appear to be φ without being φ'.

[7] Cf. 303C4-5.

ignores it. For these reasons, I choose the second reconstruction of the Visitor's grounds for claiming that present-day regimes imitate the genuine constitution: the Visitor's reason for making this claim has to do with the fact that present-day regimes appear to be constitutions while they are not constitutions.

Whichever of the two reconstructions one favours, the arguments that they attribute to the Visitor share an important trait. In both arguments, those who rule over present-day regimes rule over what are in fact imitations of the genuine constitution. However, neither argument requires that these rulers themselves take what they rule over to be imitations of the genuine constitution: as far as the arguments are concerned, the possibility remains open that those who rule over kingships, tyrannies, etc. regard them (wrongly) as genuine constitutions. To view the matter from a different angle, in the arguments attributed to the Visitor by the two reconstructions, present-day regimes are called 'imitations' because they themselves 'imitate' something, but they are not called 'imitations' because the human beings who promote and support them intentionally bring it about that they 'imitate' that something.

3. *Better and worse imitations*

In passages T1 and T2 the Visitor makes two claims: first, that all present-day regimes imitate the genuine constitution; secondly, that some present-day regimes imitate the genuine constitution for the better whereas others imitate it for the worse. The Visitor endeavours to explain or justity the second claim. He begins by saying:

T3 ΞΕ. ... ὀρθῆς ἡμῖν μόνης οὔσης ταύτης τῆς πολιτείας ἣν εἰρήκαμεν, οἶσθ' ὅτι τὰς ἄλλας δεῖ τοῖς ταύτης συγγράμμασι χρωμένας οὕτω σῴζεσθαι, δρώσας τὸ νῦν ἐπαινούμενον, καίπερ οὐκ ὀρθότατον ὄν; 297D5–D8

VIS. … given that in our view the only correct constitution is the one we have spoken about [*sc.* the one whose government is based on the statesman's knowledge, cf. 293C5-8], are you aware that the others [*sc.* regimes that are not the genuine one] must save themselves by using the written rules of this one [*sc.* the genuine constitution], by doing what is now praised, although it is not the most correct thing? (Pl. *Plt.* 297D5-8)

He immediately goes on to explain what 'what is now praised' (297D7) is:

T4 ΞΕ. Τὸ παρὰ τοὺς νόμους μηδὲν μηδένα τολμᾶν ποιεῖν 297E1
τῶν ἐν τῇ πόλει ... 297E2

VIS. That nobody in the state should dare to do anything contrary to the laws ... (Pl. *Plt.* 297E1–2)

It remains unclear what the 'written rules' (297D6) of the genuine constitution are. Earlier passages in the dialogue open up two possibilities. First, the written rules of the genuine constitution could be the laws which a statesman is obliged to issue for pragmatic reasons, namely because it is pragmatically impossible for him to evaluate and decide on the indefinitely many particular cases that could turn up (cf. 294C10-295B6). Secondly, they could be the laws which a statesman who anticipates being away for a long time has written down as reminders for his subjects (cf. 295B7-296A4). In fact, these need not be two distinct alternatives: for, the temporarily absent statesman who leaves written laws as reminders for his subjects during his absence could be a sort of 'enlargement' of the statesman who cannot pragmatically follow all the indefinitely many and varied cases that come up and therefore issues laws that will take care of the cases he cannot attend to. In this case, the laws which a statesman who anticipates being away has written down as reminders would coincide with the laws which a statesman is obliged to issue for pragmatic reasons.

4. *Mistrust of politicians*

The Visitor and Young Socrates continue their explanation of the claim that some present-day regimes imitate the genuine constitution for the better whereas others imitate it for the worse by developing (297E8-302B4) an elaborate analogy with an imaginary situation involving 'the likenesses [εἰκόνες] to which one must always compare the kingly rulers' (297E8-9), namely a steersman and a doctor[8]. It is a sort of thought-experiment where a situation is imagined in which progressively tighter restrictions are imposed on medicine and steersmanship. The restrictions are introduced in four stages.

[8] A steersman was already mentioned at 296E4-297A2, doctors at 293B1-C3, 295B10-E2, and 296B5-C3.

The first stage (297E11-298E4) is about the origin and the application of rules that codify artistic practice. Imagine a situation where someone is an exceptionally skilled doctor but the majority think that he is doing terrible things to them (I concentrate on the case of the doctor, that of the steersman is parallel): the majority think that this doctor saves only the ones he wishes to save, that he harms them for fees that he then spends not for his patients but for himself and for his own household, etc. Since the majority think this, they decide to convene a council that comprises either all the population or only the rich and contains individuals of all sorts – in particular, it does not contain only doctors but also laymen in medicine. This council issues rules about medical matters. Once these rules have been issued, they are engraved in stone and all medical practice is expected to be carried out in accordance with them. The rules have their origin in the agreement between the members of the council; but once they have been chosen, they have supreme authority. Young Socrates remarks (298E4) that such a situation would be 'very strange'.

In the second stage (298E5-10) officers that belong either to the mass of the whole population or to the group of the rich are chosen annually and are required to carry out medical practice in accordance with the written rules. The officers are chosen by lot, so there is no guarantee that they will have any medical competence. Young Socrates notes (298E10) that a situation of this sort would be 'even harder to take'.

The third stage (298E11-299B1) introduces a mechanism to examine the behaviour of the officers. At the end of each officer's yearly mandate, a court is set up whose members are either elected or chosen by lot. Thus, the judges in this court do not in general include medical experts. They are expected to examine whether the officers have operated according to the written rules. Penalties or fines may be imposed on those who are found not to have followed the rules. Young Socrates observes (299A8-B1) that whoever willingly accepted to operate as an officer in the circumstances described deserves whatever punishment is imposed on him.

In the fourth stage (299B2-E10) an additional law is introduced that forbids original and independent medical research. If anyone were to conduct research of this sort, he would not be called a doctor but a 'stargazer' and a 'babbling sophist'. Anyone would have the right to indict him and bring him before a court as corrupting the young and inducing them to practice medicine not in accordance with the laws, and if he were found guilty then the most extreme penalties would be imposed on him. The same holds for all other arts and disciplines. Young Socrates

comments (299E6-10) that in such a situation the arts would 'be completely destroyed'.

The expressions 'stargazer' ('μετεωρολόγος', 299B7) and 'babbling sophist' ('ἀδολέσχης τις σοφιστής', 299B8) make of this passage an unmistakeable allusion to the vicissitudes of (the elder) Socrates, who was attacked by means of expressions of this sort in comedy[9] and by the general public[10]. The allusion is confirmed by the mention of an indictment for corrupting young people: Socrates was indicted for corrupting the young and for not believing in the gods of the city[11]. Similar allusions occur elsewhere in Plato's dialogues[12]. However, the *Statesman*'s allusion has a novel aspect: it suggests that Socrates' condemnation by the Athenian democracy was not the result of unfortunate chance; rather, it derived from a fundamental and unavoidable incompatibility between Socrates' genuinely philosophical thought and present-day states[13].

The thought-experiment concerning medicine helps to explain the origin of laws and the way in which they are applied in present-day regimes. The citizens of present-day regimes believe that there could never be a ruler who combined the knowledge of political matters with the moral qualities that would refrain him from exercising his absolute and unchecked power for corrupt and malevolent ends (cf. 301C6-E5). A passage in Herodotus' *Histories* (3. 80) bears witness to this mistrust because it criticises monarchic rule by pointing out that if absolute power were given even to 'the best man on earth', it would corrupt him and breed arrogance. Plato himself seems in fact to share this mistrust. For, in the *Statesman* he is elusive about whether any genuine statesman actually exists or could exist[14], and in the *Laws* (9. 874E8-875D5) he is pessimistic about the possibility of any such figure ever arising (he indicates that the weakness of human nature would unavoidably entail features such as the ones feared by most people). Their mistrust of rulers prompts the citizens of present-day regimes to set up a council that consists either of the people all together or only of the rich and is supposed to issue laws, which then acquire supreme authority. Thus, laws have their origin in the agreement

[9] Cf. Ar. *Nu.* 228-230; 359-360; 1480; 1485; fr. 490 Kock; Eup. fr. 352 Kock.

[10] Cf. *Ap.* 18B7-C1; 19C2-5; 23D6-7; X. *Oec.* 11.3; *Smp.* 6.6.

[11] Cf. *Euthphr.* 2C3-3A5; *Ap.* 24B8-C1.

[12] Cf. *Phd.* 70C1-2; *Phdr.* 269E4-270A1; *R.* 6. 488E4-489A1; *Prm.* 135D3-6; *Tht.* 195B9-C4; *Sph.* 225D7-11.

[13] Cf. El Murr 2014: 249-250.

[14] Cf. Rowe 2005: 236.

between the members of the council; but once they have been issued, they have supreme authority. This account of the origin and use of laws corresponds to the first stage of the Visitor's thought-experiment[15].

In *Republic* 2 (358E3-359B5) Glaucon offers a different account of the origin of law: to inflict injustice is naturally good whereas to suffer it is naturally bad; people are unable to inflict injustice without suffering it; since the badness that comes from suffering injustice exceeds the goodness that derives from inflicting it, people conclude that it is profitable to create laws, which prevent them both from inflicting and from suffering injustice. Neither the account of the origin of law in the *Republic* nor that in the *Statesman* is to be taken as a serious attempt to offer a historically plausible reconstruction. Rather, both accounts are imaginary stories whose purpose is to clarify certain aspects of conceptions of law that the two dialogues are examining.

5. *Ignorant politicians who flout the laws*

After describing the disastrous consequences of a legal straight-jacket imposed on the arts, the Visitor and Young Socrates consider an even worse development (300A1-E3). Suppose that the officers who must exercise the arts by applying the written rules or the judges who must assess the officers' conduct were to take no notice of the written rules, either for their own profit or to do personal favours. Such a situation would be even worse than the one where the rules are respected. What corresponds to this in the case of politics is a regime whose rulers not only are ignorant in that they do not have the special form of knowledge that is statesmanship, but also take no notice of the written rules and customs and thereby put themselves above the law (and in this respect resemble the genuine statesman, whose knowledge puts him in a position to modify the laws he himself has issued)[16].

In passage T1, the Visitor remarked that 'one must say that they [*sc.* pre-

[15] Cf. El Murr 2014: 248-249.

[16] According to Griswold 1989: 156, the rulers of the degenerate case are ignorant not only because they do not have the special form of knowledge that is statesmanship, but also because they ignore that they are thus ignorant. As far as I can see, this is not required by argument in the relevant portion of the text: the rulers of the degenerate case could well be in bad faith in that they are aware of their own ignorance of statesmanship but consciously pretend to be competent in it.

sent-day regimes] ‹are› not genuine nor really being ‹constitutions›, but imitating this one [*sc.* the genuine constitution], those we speak of as well-governed for the better, the others for the worse' (293E3-5). The expression 'the others' implies that every present-day regime imitates the genuine constitution either for the better or for the worse: there are no intermediate cases, no present-day regimes that imitate the genuine constitution neither for the better nor for the worse. In the development of his argument (in the long and elaborate analogy of 297E8-302B4), the Visitor states that present-day regimes whose ignorant rulers put themselves above the regime's written laws and its customs and take no account of them imitate the genuine constitution 'utterly badly [παγκάκως]' (300E1). Degenerate present-day regimes of this sort probably coincide with those that imitate the genuine constitution 'for the worse'. On the other hand, the Visitor also suggests that present-day regimes whose ignorant rulers respect the laws imitate the genuine constitution 'finely [καλῶς]' (301A1). Such law-abiding present-day regimes probably coincide with those that imitate the genuine constitution 'for the better'. Thus, all present-day regimes are merely imitations of the genuine constitution, namely the regime based on statesmanship whose rulers issue laws only for pragmatic reasons and are free to modify these laws. However, among these present-day regimes whose status is merely that of imitations, those where the rulers respect the laws are superior to those where the rulers take no notice of the laws in order to promote their own interest or that of their friends (even though the rulers who take no notice of the laws share a trait with genuine statesmen). The present-day regimes whose rulers respect the laws are probably those that imitate the genuine constitution for the better; present-day regimes whose ignorant rulers take no notice of the laws are probably those that imitate the genuine constitution for the worse.

6. *How can the written rules of the genuine constitution be accessed?*

The Visitor says that present-day regimes 'must save themselves by using the written rules of this one [*sc.* the genuine constitution]' (297D6-7). At a later stage of the discussion he remarks that some present-day regimes are governed according to

T5 ΞΕ. ... τοὺς νόμους τοὺς ἐκ πείρας 300B1
πολλῆς κειμένους καί τινων συμβούλων ἕκαστα
χαριέντως συμβουλευσάντων καὶ πεισάντων θέσθαι τὸ
πλῆθος ... 300B4

VIS. … the laws that have been established on the basis of much experiment, with some advisers having cleverly[17] given advice on each subject and having persuaded the majority to pass them … (Pl. *Plt.* 300B1-4)

And:

T6 ΞΕ. νῦν δέ γε ὁπότε οὐκ ἔστι γιγνόμενος, ὡς δή 301D8
φαμεν, ἐν ταῖς πόλεσι βασιλεὺς οἷος ἐν σμήνεσιν ἐμφύεται, 301E1
τό τε σῶμα εὐθὺς καὶ τὴν ψυχὴν διαφέρων εἷς, δεῖ δὴ
συνελθόντας συγγράμματα γράφειν, ὡς ἔοικεν, μεταθέον-
τας τὰ τῆς ἀληθεστάτης πολιτείας ἴχνη. 301E4

VIS. But in the present situation, when – as we say – kings are not born in cities like those in beehives, single individuals straightaway superior in body and mind, it is necessary – so it seems – for people to come together and write down written rules, running after the traces of the truest constitution (Pl. *Plt.* 301D8-E4).

'The laws that have been established on the basis of much experiment, with some advisers having cleverly given advice on each subject and having persuaded the majority to pass them' (300B1-4 = T5) are probably laws that ordinary legislators of present-day regimes find by 'running after the traces of the truest constitution' (301E3-4 < T6).

One might wonder how the Visitor can consistently claim that some of the laws promulgated by ordinary legislators are laws delivered by the specific form of knowledge that is statesmanship. Doesn't such an identification generate an inconsistency? After all, one of the main messages of the part of the *Statesman* to which passage T7 belongs is that ordinary legislators lack knowledge[18].

However, on reflection, the inconsistency evaporates. The laws in question are probably (not concrete inscriptions or events or states of stating or judging, but) prescriptive propositions[19] that are the contents of cognitive states (e.g. of states of judging or knowing), of speech-acts (e.g. events of stating), and of concrete inscriptions. Just as one and the same proposition can be known by Tim and at the same time judged but not known by Jim

[17] Cf. LSJ *s.v.* 'χαρίεις' III 1. The adverb could also be rendered by 'in an attractive way', in which case it would be indicating that the advisers presented the laws to their respective assemblies in a convincing way.

[18] Cf. Rowe 1995a: 16-17; Rowe 1999: XV.

[19] Prescriptive propositions are not acknowledged by mainstream modern philosophical logic but were accepted by ancient Stoic logic (cf. D.L. 7.67; S.E. *M.* 8.71).

and doubted by Frank, so also one and the same law, a prescriptive proposition, can be issued by knowledgeable statesmen and also be promulgated by ordinary legislators, who have discovered it 'on the basis of much experiment' (300B1-2) and have profited of the advice of some advisers who have also persuaded the majority. In some cases, ordinary legislators, on the basis of experience and some expert advice, happen to light on laws (prescriptive propositions) which are also delivered by statesmanship. When this happens, ordinary legislators of present-day regimes do not have knowledge of these laws because only genuine statesmen have knowledge about political matters and ordinary legislators of present-day states are not genuine statesmen. Just as ignorant individuals can make true judgements without having knowledge of what they truly judge, so also ordinary legislators of present-day regimes can find some of the best possible laws, i.e. some of the laws that a genuine statesman issues, or would issue, for pragmatic reasons, and they can do this without having the statesman's knowledge[20]. The parallel between the distinction between true judgement and knowledge, on the one hand, and the distinction between the laws found by regimes 'on the basis of much experiment, with some advisers having cleverly given advice on each subject and having persuaded the majority to pass them' (300B1-4), and the laws issued by a genuine statesman for pragmatic reasons, on the other, is confirmed by 301B2-3: here the Visitor refers to the genuine king, i.e. the statesman, and the king of a normal law-abiding monarchy by means of the phrase 'the one who rules on his own according to laws with knowledge or with judgement [τὸν μετ' ἐπιστήμης ἢ δόξης κατὰ νόμους μοναρχοῦντα]'. Thus, while the laws of the genuine king are issued on the basis of knowledge, the laws of the king of a normal law-abiding monarchy are issued on the basis not of knowledge but of mere judgement.

7. *The second application of the concept of imitation: laws*

In a difficult and variously interpreted passage, the Visitor mentions again the concept of imitation:

T7 ΞΕ. Οὐκοῦν μιμήματα μὲν ἂν ἑκάστων ταῦτα εἴη τῆς ἀληθείας, τὰ παρὰ τῶν εἰδότων εἰς δύναμιν εἶναι γεγραμμένα; 300C5 300C7

[20] Cf. C. Gill 1995: 296; Hirsch 1995: 186; Palumbo 1995: 180; Márquez 2012: 277; El Murr 2014: 252-253.

VIS. Wouldn't[21] then these be imitations of the truth of each and every thing, things written down so far as possible by those who know? (Pl. *Plt.* 300C5-7)

Passage T7 raises several exegetical problems. I concentrate on two.

The first exegetical problem that I intend to discuss concerns the relationship between the expressions 'παρὰ τῶν εἰδότων' (300C6) and 'γεγραμμένα' (300C6-7). One possibility is that these two expressions could be reciprocally independent ('things issuing from those who know that have been written down so far as possible'); another possibility is that the first expression could be the complement of agent for the second ('things written down so far as possible by those who know'). Since in the presence of a verb in the passive a phrase consisting of 'παρά' followed by the genitive is most naturally understood as a complement of agent[22], the second solution is more likely[23]. In this case, the words 'εἰς δύναμιν εἶναι' (a single adverbial phrase[24], 'so far as possible') at 300C6 probably modify the whole of the rest of the phrase in which they are embedded: what is described as being the case 'so far as possible' is that the things in question should have been 'written down […] by those who know'[25]. Note that the extent to which the things in question have been 'written down […] by those who know' could well be minimal.

The second exegetical problem that I want to consider concerns the occurrence of 'these' ('ταῦτα') at 300C5. One possibility is that it could refer forward, so as to create an antecedent for the explication given in the

[21] For the use of an isolated 'μέν' in rhetorical questions, cf. LSJ *s.v.* 'μέν' A I 3. The combination of 'οὐκοῦν' with an isolated 'μέν' in a rhetorical question is common in Plato: cf. *Cra.* 407C6-7; *Tht.* 210B8-9; *Sph.* 265A4-5; *Plt.* 278C3-6; etc.

[22] Cf. LSJ *s.v.* 'παρά' A II 4; Smyth 1920, 371; Pl. *Phdr.* 245C1; *R.* 4. 499D5-6.

[23] Cf. Stallbaum 1841: 289; Giorgini 2005: 325.

[24] Cf. Stallbaum 1841: 289; Campbell 1867: *Plt.* 157; Rowe 1995a: 231.

[25] Had one chosen the first alternative, i.e. treating the expressions 'παρὰ τῶν εἰδότων' (300c6) and 'γεγραμμένα' (300c6-7) as reciprocally independent, the further problem would have arisen of deciding what the adverbial phrase 'εἰς δύναμιν εἶναι' modifies: it could have modified either 'παρὰ τῶν εἰδότων' ('those issuing so far as possible from those who know that have been written down') (cf. Skemp 1952: 209; Rowe 1995c: 27; Márquez 2012: 269, 279), or 'εἰδότων' ('those issuing from those who know so far as possible that have been written down') (cf. Fowler and Lamb 1925: 155; Lane 1995: 287), or 'γεγραμμένα' ('those issuing from those who know that have been written down so far as possible') (cf. Stallbaum 1841: 289; Jowett 1892: IV 504; Taylor 1961: 324; Warrington 1961: 280; Adorno 1988: I 945; Annas and Waterfield 1995: 68). Fraccaroli 1911: 308 takes the occurrence of 'εἰς δύναμιν εἶναι' at 300C6 to modify that of 'μιμήματα' at 300C5, but this is grammatically impossible. Teisserenc 2005: 377 takes 'εἰς δύναμιν' to modify 'εἶναι', but this is also grammatically impossible.

passage's second half ('These are the things that would be imitations … – namely things written down …')[26]; alternatively, it could refer backwards, so as to pick up the occurrence of 'laws and written rules' ('νόμους καὶ συγγράμματα') at 300C1-2, in the Visitor's remark that immediately precedes the one that constitutes T7[27]. In principle, the occurrence of 'these' ('ταῦτα') at 300C5 could refer forward; but the shortly preceding occurrence of 'these' ('ταῦτα') at 300C2 (cf. also its occurrence at 300B4) speaks in favour of the second alternative, according to which it refers backwards[28].

I thus take it that the claim made in passage T7 is that all laws of present-day regimes are 'imitations of the truth of each and every thing' (300C5-6). This claim seems to be presented as an inference (cf. the occurrence of 'then', 'οὐκοῦν', at 300C5), and the reason justifying this inference is probably given in T7's second half: the reason why all laws of present-day regimes have the status of 'imitations of the truth of each and every thing' (300C5-6) is that they are 'things written down so far as possible by those who know' (300C6-7). In the context of the argument of this part of the *Statesman*, a form of 'to imitate φ' may be plausibly taken to introduce an idea that could also be conveyed by the corresponding form of 'to appear to be φ without being φ' or 'to instil the illusion of being φ'[29]. It may therefore be plausibly inferred that passage T7 is providing some justification for the view that all laws of present-day regimes appear to be the truth without being the truth. Since the truth in question is probably the specific form of knowledge that is statesmanship (cf. 300D10)[30], the thesis put forward in passage T7 is probably that all laws of present-day regimes appear to be the specific form of knowledge that is statesmanship without being such a thing. The reason why they have this appearance is that they have some link with knowledge, in particular with statesmanship: for they are 'things written down so far as possible by those who know' (300C6-7), and this is because they have been found 'with some advisers having cleverly given advice' (300B2-3)[31]: in the extended analogy developed by the Visitor in the preceding pages, the 'advisers' who have given advice are the few knowledgeable experts who together with laymen form the committees that

[26] Cf. Rowe 1995a: 16-17, 230-231; Rowe 1995c: 26-27; Rowe 1997: 278-279; Rowe 1999: XV-XVI; Rowe 2001: 71-72; El Murr 2014: 253.

[27] Cf. Jowett 1892: IV 504; Fowler and Lamb 1925: 155; Annas and Waterfield 1995: 68.

[28] Teisserenc 2005: 378 also believes that 'these' ('ταῦτα') at 300C5 should refer backwards.

[29] Cf. above, nn. 5 and 7 and the paragraphs to which they are appended.

[30] Cf. Lane 1995: 287; Palumbo 1995: 181.

[31] Cf. Campbell 1867: *Plt.* 157.

issue the laws (cf. 297E11-298E4, esp. 298D5-7, where the Visitor speaks of 'some doctors and steersmen giving their advice [συμβουλευόντων] together with laymen'). Perhaps the fact that the laws of present-day regimes 'have been established on the basis of much experiment' (300B1-2) also contributes to their appearing (without being) the specific form of knowledge that is statesmanship. So: although the laws of present-day are not the specific form of knowledge that is statesmanship, they appear to be statesmanship, and at least part of the reason why they have this appearance is that it is generally known that some of the advisers who have contributed to issuing them have the relevant form of knowledge.

It might be objected that laws cannot appear to be statesmanship because laws and statesmanship are entities that belong to different categories: laws are prescriptive propositions, statesmanship is a mental state, and a set of prescriptive propositions obviously is not a mental state and therefore cannot appear to be a mental state. The most plausible reply to this objection is that the categorial distinction between laws as prescriptive propositions and statesmanship as a mental state is not obvious to those who fall prey to the appearance, people who are not so clear about categorial distinctions. Thus, even if laws and statesmanship are entities that belong to different categories, a set of laws may well appear to be statesmanship. Moreover, in Greek, 'τέχνη' ('art') may be used for mental states as well as for sets of rules and even treatises[32]: the people to whom the laws appear to be statesmanship are perhaps thinking of the art of statesmanship as a set of rules. Also note the remark, attributed to law itself in the extended analogy of the preceding pages, that 'nothing can be wiser [σοφώτερον] than the laws' (299C5-6), a remark that echoes a formula which Thucydides (3.37, 4) puts in the mouth of Cleon, the great democratic leader of Athens[33].

8. *Law in the* Statesman

Plato's attitude to law in the *Statesman* is complex and nuanced. On the one hand, Plato has a negative attitude to law in that he maintains that laws are too simple to cater for all the complexities of human life (cf. 293E7-294C9). On the other hand, he has a positive attitude to law in that

[32] Cf. LSJ *s.v.* 'τέχνη' II, III, and VI.
[33] Cf. Teisserenc 2005: 373.

he maintains that law is indispensable in all cases – in the case of the genuine constitution as well as in that in present-day regimes (which, properly speaking, are not constitutions). For, in the genuine constitution, which is ruled by a single genuine statesman or by a small group of genuine statesmen, law is indispensable for pragmatic reasons: although the genuine statesman or statesmen would be capable to decide about each individual case without creating or following laws, laws are needed because the citizens are too many and their cases too varied for the statesman or statesmen to be in a position to decide about each individual case that could come up (just as expert gymnastics trainers are obliged to prescribe shared diets and shared exercises to groups of trainees because it is practically impossible for them to set out personalized diets and personalized exercises) (cf. 294C10-295B6). As for present-day regimes, laws are necessary because their rulers lack the genuine statesman's knowledge and are therefore not competent to decide about individual cases without laws to which to attend (cf. 297D4-E5).

Although law is necessary in all cases, there remains a difference between law in the genuine constitution and in present-day regimes. In a genuine constitution, the genuine statesman should override or modify laws when his knowledge tells him that he should (cf. 295B7-296A4). In present-day regimes, the rulers should never override or change laws because their lack of knowledge would probably lead them to disastrous modifications. Thus, law is changeable in the genuine constitution but should be unchangeable in present-day regimes. All present-day regimes imitate the genuine constitution, i.e. appear to be genuine constitutions without being such a thing: those that do respect their laws imitate the genuine constitution 'for the better' in that the amount of harm they inflict on the citizens is somehow limited; those that fail to respect their laws imitate the genuine constitution 'for the worse' in that they foster an extremely unhappy life of the state.

The laws of present-day regimes have a very tenuous link with the genuine statesman's knowledge: the link consists in the fact that alongside many laymen, some statesmen have also contributed their advice with a view to issuing these laws. This tenuous link suffices to give laws the appearance of being knowledge, in particular the specific form of knowledge that is statesmanship. Since they appear to be statesmanship without being such a thing, the laws of present-day regimes may be described as 'imitations of the truth' (300C5-6), i.e. of statesmanship. In some lucky cases, some of the laws issued by the ignorant rulers of a present-day regime are precisely those which are issued by genuine statesmen. There is no incon-

sistency here because laws are prescriptive propositions and the same prescriptive proposition can be issued both by an ignorant ruler of a present-day regime and by a genuine statesman. Even in such a case, the epistemic attitude which an ignorant ruler and the genuine statesman have to one and the same prescriptive proposition are different: only the genuine statesman has knowledge about that prescriptive proposition, the ignorant ruler only makes a judgement regarding it.

9. *Did Plato change his mind about the merits of democracy?*

Some commentators believe that the *Statesman* commits Plato to a revaluation of democracy and of the role of law with respect to the position presented in the *Republic*, where democracy was described as the last step before the catastrophy of tyranny and the rule of the state was entrusted to philosopher-kings. Some even suspect that a justification of the condemnation of Socrates by the Athenian democracy is in the offing. For, the Visitor seems first to describe the condemnation of Socrates as a consequence of the supreme authority of law in society and then to claim that respecting the laws is the best possible course of action for present-day regimes, where no genuine statesman is in a position of power[34].

However, the conclusions about the revaluation of democracy and the justification of the condemnation of Socrates cannot be safely drawn. For, the regimes where law has supreme authority are consistently described in the *Statesman* as surrogates of the only true constitution, where statesmanship is at the helm, surrogates whose widespread occurrence is due to the commonly held view that no statesman could ever be above the temptations that come with absolute power. Moreover, the various strictures which the Visitor describes as consequences of the majority's mistrust of politicians are at least in part gratuitous and belong to a caricature. In particular, even if it is granted that laws should have supreme authority, it does not follow that philosophical inquiry about ethical and political matters, and in particular about justice and the value of law, should be forbidden. Socrates' own life, as it is described in the *Crito* (especially at 50A6-54E2), shows that the absolute respect of the laws is compatible with free philosophical inquiry about ethical and political matters. A veto on philosophical inquiry is unjustified even in a state where law enjoys supreme authority.

[34] Cf. Sabine 1962: 74; Griswold 1989: 157-162; Annas and Waterfield 1995: XVIII-XX.

10. *A new classification of forms of government*

The distinction between regimes that imitate for the better and for the worse yields a total number of seven forms of government: (1) the genuine constitution, where the ruler is the authentic statesman who governs on the basis of statesmanship; (2) the monarchy based on laws that the monarch respects (kingship); (3) the monarchy based on laws that the monarch ignores to pursue his personal interest (tyranny); (4) the government of the few based on laws that the rulers respect (aristocracy); (5) the government of the few based on laws that the rulers ignore to pursue their personal interests (oligarchy); (6) the government of the many based on laws that the rulers respect (democracy); (7) the government of the many based on laws that the rulers ignore to pursue their personal interests (democracy). Earlier (at 291C9-292A4) only five types of regime had been distinguished because the two subdivisions of democracy had not been distinguished (the two inquirers had relied on linguistic usage, which has a single name, 'democracy', for both subdivisions) and the regime where statesmanship is at the helm had not yet been isolated. The noun 'democracy' is used both for the government of the many where the laws are respected and for that where the laws are ignored. Similarly, the noun 'king' is used both for the monarchic ruler who governs on the basis of the statesman's knowledge and for the monarchic ruler who relies on laws that he respects. This enables the Visitor to draw a conclusion that sounds enigmatic: 'As a result of this the five names of what are now called constitutions have become only one' (301B7-8). Some commentators find this remark so strange that they emend the text (David Robinson transposes a modified form of it to 301C7)[35]. But the text of the MSS may be defended: there is only one name of constitutions because really there is only one constitution (the others are only imitations). The name of the only constitution is 'kingship', a name it shares with one of the imitations.

The preceding considerations show that we should not wonder at the evils that afflict present-day states. Rather, we should wonder at the fact that despite their shortcomings, many present-day states survive (though some of them 'sink like ships and perish', 302A6-7). This negative evaluation of present-day states and the call for an enlightened rule echo similar remarks in the *Republic* (cf. 5. 473B4-E5).

[35] Cf. D.B. Robinson 1995: 41.

11. *The quality of life in the various regimes*

The Visitor then offers a ranking of the imitative regimes where it is harder or easier to live. All the law-abiding regimes are easier to live in than the law-flouting ones. When the regime is law-abiding, the one where it is easiest to live is the monarchy, followed by the government of the few, followed by the government of the many. By contrast, in the case of the law-flouting regimes, the ranking is reversed: the one where it is easiest to live is the government of the many, followed by the government of the few, followed by the monarchy (which amounts to tyranny).

The Visitor does not explain why democracy is 'the worst of the best and the best of the worst'. The most plausible explanation is that the fragmentation of power that is typical of it makes it weak and therefore unable to give rise to anything great either among good things or among bad ones: when a democracy is law-abiding, respect of law in it will be less efficient than in a law-abiding monarchy (kingship) or in a law-abiding government of the few (aristocracy) (in a law-abiding democracy, the harmonious operation of a large group of people is slower because the large number of offices requires a multiplication of laws and procedures); when a democracy is law-flouting, its lack of respect for the law will be of less consequence than in a law-flouting government of the few (oligarchy) or a law-flouting monarchy (tyranny) (in a law-flouting democracy, the conflicting interests of the many will to some extent cancel each other out and reduce the total damage).

12. *The third application of the concept of imitation: present-day politicians*

The Visitor and Young Socrates state that present-day politicians should be separated from the genuine statesman and may be described as sophists:

T8 ΞΕ. Οὐκοῦν δὴ καὶ τοὺς κοινωνοὺς τούτων τῶν 303B8
πολιτειῶν πασῶν πλὴν τῆς ἐπιστήμονος ἀφαιρετέον ὡς 303C1
οὐκ ὄντας πολιτικοὺς ἀλλὰ στασιαστικούς, καὶ εἰδώλων
μεγίστων προστάτας ὄντας καὶ αὐτοὺς εἶναι τοιούτους,
μεγίστους δὲ ὄντας μιμητὰς καὶ γόητας μεγίστους γίγνε-
σθαι τῶν σοφιστῶν σοφιστάς. C5

ΝΕ. ΣΩ. Κινδυνεύει τοῦτο εἰς τοὺς πολιτικοὺς λεγο-
μένους περιεστράφθαι τὸ ῥῆμα ὀρθότατα. 303C7

VIS. We must therefore remove also those who participate in all these constitutions, except for the one based on knowledge, as being, not statesmen, but factious, and ‹we must say› that by being rulers of the greatest images they themselves also are such, and that by being the greatest imitators and beguilers they turn out to be the greatest sophists among sophists[36].

Y.S. This expression [*sc.* 'sophist'] may happen to have been only too correctly turned round against the so-called statesmen (Pl. *Plt.* 303B8-C7).

In passage T8 the Visitor asserts that 'we must [...] remove' (303C1) present-day politicians because they are factious. The removal in question is probably not a 'physical' removal such as exile or assassination, but a 'logical' removal: it is the setting apart of present-day politicians from genuine statesmen[37].

The description of all present-day politicians as 'factious' (303C2) is a bit surprising: do law-abiding rulers deserve it? In a passage of the *Laws* (8. 832B10-C3), the Athenian claims that tyranny, oligarchy, and democracy are not properly speaking 'constitutions' but 'factious systems' because in these regimes the rulers 'never hold power with the consent of the governed' (832C3-4). It is difficult to undersand on what grounds the Athenian can claim that a democracy fails to have the consent of the governed. Several explanations of the *Statesman*'s description of present-day politicians as factious are possible. One possibility is that Plato has been carried away partly by rhetoric and partly by his low opinion of present-day politicians in Athens, and has therefore been led to draw an invalid conclusion. Alternatively, he might be implicitly restricting his consideration to the rulers of law-flouting constitutions (i.e. tyrannies, oligarchies, and law-flouting democracies), who ignore the laws because they are driven by ambition and the desire for power and therefore give rise to factions[38]. Yet

[36] There is a minor puzzle concerning the syntax of 303C2-5: what governs the infinitives 'εἶναι' (303C3) and 'γίγνεσθαι' (303C4-5)? One possibility is to supply an understood 'ὥστε' (immediately before the 'καί' at 303C2) (cf. Fischer 1774: 187). Another is to regard the infinitives 'εἶναι' and 'γίγνεσθαι' as governed by the 'ὡς' of 303C1, which acquires a declarative sense (cf. Stallbaum 1841: 298). A further possibility is to supply an understood 'λεκτέον' (immediately after the 'καί' at 303C2) that functions as the main verb governing the following infinitives (cf. Stephanus 1578: II 303; Rowe 1995a: 236). In my translation I adopted the last solution: the words 'we must say' render the understood 'λεκτέον').

[37] Cf. Rowe 1995a: 236. For a similar 'logical' use of 'to remove' ('ἀφαιρεῖν'), cf. 262B1; D3; 263C9; 268E1; 291C6; 292D6.

[38] Cf. Rowe 1995a: 236.

another possibility is that Plato could be attributing to the Visitor the view that the lack of knowledge that characterizes all present-day politicians, i.e. their failure to master statesmanship, and the fact that the laws on which present-day regimes are based merely appear to be statesmanship inevitably bring it about that the states ruled by present-day politicians will sooner or later be torn apart by factions[39].

At the beginning of his examination of present-day politicians, the Visitor playfully described them in terms that recall the characters of a satyr play (cf. 291A8-B2)[40]. He resorts to this light-hearted description again at the end of his examination (cf. 303C8-D2). Passage T8 provides the key for understanding the joke: present-day politicians are imitators, and drama is the realm of imitation (cf. 288C2-3).

13. *The Visitor's argument*

Passage T8 contains a brief argument for the thesis that present-day politicians are the greatest of sophists. The premiss of this argument is that present-day politicians are 'rulers' (303C3) of regimes that are 'images [εἴδωλα]' (303C2) of the genuine constitution, the constitution 'based on knowledge' (303C1). Earlier (at 293E3-6 = T1 and 297C1-4 = T2) the two inquirers had agreed that present-day regimes are 'imitating [μεμιμημένας]' (293E4) the genuine constitution and are 'imitations [μιμήματα]' (297C2) of it. It may be plausibly assumed that the nouns 'image' ('εἴδωλον') and 'imitation' ('μίμημα') are mere stylistic variants: this assumption is confirmed by other occurrences of the two nouns in the *Statesman* (at 306D2 and D3), by how they are used in the *Sophist*[41], and by how the names of the corresponding crafts, 'εἰδωλοποιικήʼ and 'μιμητικήʼ, are used in the *Sophist*[42]. In view of this, the earlier agreement that present-day regimes, which are of course the regimes of which present-day politicians are rulers, are 'imitations [μιμήματα]' (297C2) of the genuine constitution amounts to an endorsement of the premiss of T8's argument, that present-day politicians are 'rulers' (303C3) of regimes that are 'images [εἴδωλα]' (303C2) of the genuine constitution. 'By being rulers of the greatest images' (303C2-3), present-day

[39] Cf. Márquez 2012: 295.
[40] On this comparison, cf. El Murr 2014: 221-223.
[41] Cf. 234B6 with C5; Bondeson 1972: 1.
[42] Cf. 235B8-9; 235C3; 235D1-2; 236C6-7; 265B1-2; Kamlah 1963: 28.

politicians are themselves 'such' (303C3), namely 'the greatest images', and are therefore 'the greatest imitators' (303C4). Since to be a sophist is to be an imitator, present-day politicians are 'the greatest sophists among sophists' (303C4-5). The images of which present-day politicians are 'rulers' (303C3), namely present-day regimes, are 'the greatest' because of their importance[43]. The argument is qualified throughout by the adjective 'greatest' because of the importance of the images, i.e. regimes, of which present-day politicians are rulers. It vindicates the correctness of the earlier description of the present-day politician as 'the greatest beguiler of all the sophists and the most expert in their art' (291C3-4).

The inference from the claim that present-day politicians are 'rulers of the greatest images' to the claim that they are themselves 'such', namely 'the greatest images', is probably based on the thought that since they are rulers of regimes that are images of the genuine constitution, present-day politicians are themselves images of the ruler of the genuine constitution, namely of the genuine statesman[44]. This matches an earlier remark by the Visitor to the effect that present-day politicians 'pretend to be statesmen and convince many [*sc.* that they are statesmen], but are not [*sc.* statesmen] in any way at all' (292D6-8, cf. 293C7-8)[45]. Since they are images, or imitations, of the genuine statesman, present-day politicians may be described as imitating the genuine statesman[46], and therefore as imitators of him (they are like mimes, who imitate by appearing to be people or things that they are not)[47].

14. *The* Sophist *on sophists, images, and falsehood*

The *Statesman*'s description of the present-day politician as 'the greatest beguiler of all the sophists and the most expert in their art' (291C3-4) involves a cross-reference to the *Sophist*, where the sophist was described as 'a beguiler and an imitator' (235A8, cf. 235A1; 241B6-7). Since the way

[43] For the use of 'great' ('μέγας') to express importance, cf. LSJ *s.v.* 'μέγας' A II 4.

[44] In the *Apology* (20C6-7) Socrates narrates that he came to the view that present-day politicians seem (to many and to themselves) to be wise without being wise. Although he does not use the concept of imitation, recall the connection between 'to imitate φ' and 'to appear to be φ without being φ' (cf. above, n. 5 and text thereto).

[45] Cf. Palumbo 1995: 178-179.

[46] Recall that in T1 the verb 'to imitate' ('μιμεῖσθαι', 293E4) describes the imitations, not their authors.

[47] Cf. *R.* 3. 393C5-6; *Sph.* 267A1-B3.

in which the themes of images, imitation, and appearance are developed in the *Sophist* is likely to shed some light on T8's argument for the thesis that present-day politicians are the greatest sophists, I shall examine the *Sophist*'s treatment of these themes.

In the *Sophist*, the Visitor and Theaetetus try to define the sophist by using the method of division. After practicing the method by applying it to an easier case that serves as a model, that of the angler (218C5-221C5), they direct it to the sophist and obtain six accounts of him (221C6-231B9): the sophist is (1) a hunter of rich and prominent young men (221C6-223B7, 231D2-4), (2) a seller of speeches and learning who buys his goods and operates in more than one city (223C1-224D3, 231D5-7), (3) a seller of speeches and learning who buys his goods and operates within a single city (224D4-E5, 231D8-10), (4) a seller of speeches and learning who produces his goods himself and operates within a single city (224D4-E5, 231D10-12)[48], (5) a verbal fighter (224E6-226A5, 231D12-E3), and (6) an educator who by means of refutation purifies the soul from its pretence of knowledge (226A6-231B9, 231E4-7). Faced with these six accounts, Theaetetus confesses: 'I am puzzled [ἀπορῶ]' (231B9). He reports that his puzzlement is due to 'the fact that the sophist has appeared in many ways' (231B9-C1). So, a new attempt is deemed necessary. The novel approach will eventually lead to a seventh account of the sophist, which is presented in the last part of the dialogue and is deemed successful[49].

15. *Appearance is of the essence*

In their comments on the first six accounts (231B9-232A7), the Visitor and Theaetetus remark several times that a sophist *appears* to have certain competences (the concept of appearance is expressed by the verbs 'φαίνεσθαι' and 'ἀναφαίνεσθαι': cf. 231B9-C1; D2; D9; 232A1-2). These remarks provide the starting point for a fresh discussion of the sophist (232B1-236D4), a discussion that aims to provide some background for the new account of him. This new account turns upon the concepts of appearing

[48] In the summary at 231D10-12 the requirement that the seller of speeches and learning who produces his goods himself should operate within a single city is dropped.

[49] Most commentators hold that the first six accounts are not successful. But there is disagreement about the seventh: some commentators (e.g. Cornford 1935: 187; Pellegrin 1991: 410; Notomi 1999: 296; M.L. Gill 2010: 184; Rickless 2010: 289, 293) maintain that it ranks as successful, others (e.g. Ryle 1966: 139; Brown 2010: 152-153, 160-163) that it also fails.

and seeming: the *essence* of the sophist is exactly his *appearing* (φαίνεσθαι) or *seeming* (δοκεῖν) to have skills and knowledges which he in fact lacks[50] (the verbs 'φαίνεσθαι' and 'δοκεῖν' are used as equivalent variants)[51]. In this respect the first six accounts, despite their failure, pave the way for the seventh, successful account.

In the discussion that prefaces their fresh attempt to define the sophist, the two inquirers hark back especially to the division leading to the fifth of the sophist's six accounts, that according to which the sophist is a verbal fighter. A sophist is a disputer (ἀντιλογικός) (232B6-7). He also teaches others to be disputers (232B8-10)[52]. He claims to do this about all subjects: he claims to make his pupils disputers about divine things hidden from common eyes, perceptible objects both in the heavens and on earth, problems of being and becoming, issues of law and politics, and questions concerning the crafts (232B11-E5). In the discussion's next step (232E6-233D2) the idea of apparent knowledge is introduced. Nobody knows everything. Sophists therefore do not know all the subjects about which they claim to teach others to become disputers. On the other hand, they bring the young to judge that 'they are the wisest of all about all things' (233B2). The reason why they do this is that 'if they did not dispute correctly nor appear [ἐφαίνοντο] to them [*sc.* to the young] to do so, and if while appearing [φαινόμενοι] to do so they did not all the more seem [ἐδόκουν] to be wise in virtue of their controversies, then […] one would hardly be willing to become a pupil of these people by giving them money' (233B3-7). Hence sophists 'appear [φαίνονται] […] to be wise about all things […] while not being so' (233C6-9). It's 'because he is an imitator of the wise man' (268C1), in Greek 'σοφός', that the sophist has a name derived from his, in Greek 'σοφιστής'[53]. It is worth pointing out that the Visitor's claim that sophists 'appear […] to be wise about all things' (233C6) is confirmed by independent evidence: in striking contrast with Socrates' disavowal of knowledge, the sophists did feign universal knowledge[54].

[50] Cf. Bluck 1963: 58; Pippin 1979: 190-191; Ledesma 2009: 237-238; Rickless 2010: 296-297; Long 2013: 128-129.

[51] Cf. 216C4-5 with C7, D1, and D2; 233B3-4 with 235A2; 233B4; 233C1 with C6; 236E1; 267C5 with C8; Vernant 1975: 128; Notomi 1999: 168; Barnouw 2002: 31.

[52] Cf. *Prt.* 312D5-7.

[53] Cf. Zadro 1961: 95; Notomi 1999: 119-121.

[54] Cf. *Euthd.* 271C5-7; 293E5-295A9; 295E4-296D4; *Prt.* 315C5-7; 315E7-316A1; *Grg.* 447C5-8; 462A8-10; *Men.* 70B5-C3; *Hp.Mi.* 363C7-D4; 368B2-E1; *Dissoi Logoi* 8. 1-13; Apelt 1897: 102-103; Cornford 1935: 191-192; Wolff 1991: 24-25; Cordero 1993: 225; Napolitano Valditara 2007: 163, 198.

16. *The sophist's imitation*

How do the sophists carry out their imitation? How do they achieve the goal of appearing to be wise about all things? The Visitor appeals to an analogy with a model (παράδειγμα, 233D3) that focuses on a graphic imitator, viz. a painter. A painter produces painted imitations (paintings) of everything and can conceal from 'those young children who are silly' (234B8) and are viewing his imitations 'from far away' (234B8) that he is cheating them into judging that he can actually produce whatever he wants (i.e. that he is a sort of god). Analogously, a verbal imitator, viz. a sophist, produces 'spoken images [εἴδωλα λεγόμενα]' (234C5-6) (statements) and can lead 'the youths who are still far from the truth about things' (234C3-4) to 'judge [δοκεῖν] that truths are being stated [ἀληθῆ λέγεσθαι] and the speaker is therefore the wisest of all about all things [τὸν λέγοντα δὴ σοφώτατον πάντων ἅπαντ' εἶναι]' (234C6-7, cf. 233B1-2). The way in which the analogy is set up suggests that just as the painter can delude the silly children into thinking that his painted imitations are what they imitate, namely people, animals, fruits, or whatever (while concealing from them that they are being deluded), so also the sophist can delude the inexperienced youths into thinking that the spoken imitations (statements) he utters are what they imitate, namely truths (while concealing from them that they are being deluded) (recall[55] that in this part of the *Sophist* the nouns 'image' and 'imitation', 'εἴδωλον' and 'μίμημα', are mere stylistic variants). Mark that the delusion caused by the verbal imitator is one whereby the inexperienced youths judge that 'truths are being stated' (234C6, cf. *R.* 2. 382D2-3): it is because they are led to judge that the sophist's statements about any subject are truths that the youths judge him to be wise about all things (producing true statements about a certain subject is an indication of 'wisdom' about that subject)[56].

Some of the ideas involved in the analogy return elsewhere in Plato's dialogues. The idea that an imitative artist produces everything returns in *Republic* 10: cf. 596B12-E11[57]; 598B6-D6. The idea that imitations can deceive their viewers or hearers to take them to be what they imitate may be found in various points of the dialogues: cf. *Sph.* 264D5-7; *R.* 3. 393A3-B2;

[55] Cf. above, n. 41 and text thereto.

[56] Cf. Notomi 1999: 134; Palumbo 2013: 273.

[57] When, a 596D1, the imitative artist who produces everything is described as a 'sophist', the expression is used in the traditional sense of 'expert' or 'master craftsman' but also alludes to the category represented by Protagoras etc.: cf. Notomi 2011: 315.

7. 523B5-6; 10. 598C1-4; 598D1-5; 600E7-601A2. In the *Gorgias* Socrates describes rhetoric in a way that recalls what the present *Sophist* passage says about the sophist's art: rhetoric does not know about the matters it deals with but 'has discovered some device of persuasion so as to appear to those who do not know that it knows more than those who do know' (459B8-C2).

One of the roles of the Visitor's analogy is to bring to the forefront the art of producing and the art of producing imitations, under which the sophist's art will eventually be subsumed (in the first five definitions it had been subsumed under the other main species of the genus art, i.e. the art of acquisition) (cf. 219A8-C1; 265A4-B1)[58]. It should not escape notice that the move is not justified by an argument nor by anything that came before: we had reached the result that the sophists' art enables them to 'appear […] to be wise about all things […] while not being so' (233C6-8); we were given an example of another art that endows its possessors a with an apparent capacity concerning all things, namely the art of producing imitations, an art whose masters appear to produce everything; now we are introduced to the idea that the way in which sophists manage to appear to be wise about all things relies on the production of imitations – specifically, spoken imitations of true statements about anything.

The inference in passage T8 recalls the one that justifies the description of the sophist as someone who appears to have universal knowledge[59]. The reason why the sophist appears to have universal knowledge is that he produces imitations of true statements in all areas. In the case of present-day politicians as well as in that of sophists, the individual's pretence to be what he is not is based on his bearing a certain relation (ruling in the case of present-day politicians, uttering in the case of sophists) to entities (regimes or statements) that are imitations of those (constitutions or true statements) to which the character whom the individual pretends to be (a statesman or an omniscient sage) bears that same relation (ruling or uttering).

17. *The Visitor's argument in passage T8*

Is the argument offered by the Visitor in T8 valid? A valid argument acceptably close to the Visitor's is the following:

[58] Cf. Notomi 2011: 321.
[59] Cf. Long 2013: 128.

[1] Every present-day politician rules over at least one regime that appears to be a constitution but is not a constitution, and every present-day politician rules only over regimes that appear to be constitutions but are not constitutions.
[2] Whoever rules over at least one regime that appears to be a constitution appears to be a statesman.
[3] Whoever rules only over regimes that are not constitutions is not a statesman.
[4] Every present-day politician rules over at least one regime that appears to be a constitution.
[5] Every present-day politician rules only over regimes that are not constitutions.
[6] Every present-day politician appears to be a statesman.
[7] Every present-day politician is not a statesman.
[8] Every present-day politician appears to be a statesman but is not a statesman.

Propositions [1]-[3] are the argument's premisses, [4]-[7] are intermediate steps, and [8] is the conclusion. Proposition [1] is a reasonable paraphrase of the claim that present-day politicians rule over regimes that are images of, or imitate, constitutions, i.e. regimes that appear to be constitutions but are not constitutions (I am assuming that forms of 'to imitate φ' introduce an idea that could be properly expressed by the corresponding forms of 'to appear to be φ without being φ' or 'to instil the illusion of being φ')[60]. Proposition [3] is a tacit assumption and is uncontroversial (at least if one ignores the case of private individuals who give competent advice to 'professional' statesmen, cf. 259A6-9). Propositions [4] and [5] follow from [1] by first-order logic. Similarly, proposition [6] follows from [4] and [2] by first-order logic. Again, proposition [7] follows from [5] and [3] by first-order logic. The conclusion [8], which follows from [6] and [7] by first-order logic, is a paraphrase of the claim that present-day politicians are images of, or imitate, statesmen.

I have not yet discussed proposition [2], which is a tacit assumption. It is controversial. Whatever plausibility it has derives from the claim that whoever rules over at least one regime that appears to be a constitution appears to rule over at least one constitution, namely to be a statesman (ruling over at least one constitution and being a statesman are treated as equivalent in the present context). This claim is objectionable: for, a certain regime could appear to be a constitution without anyone who as a matter of fact rules over it appearing to rule over it or to rule over a constitution (the actual rulers of a regime that appears to be a constitution could well be hidden). The claim can only be defended by making two assump-

[60] Cf. above, n. 5 and text thereto.

tions: first, that if a regime appears to be a constitution, then the activity of ruling of the ruler or rulers is evident to the subjects to whom the appearing pertains; secondly, that appearance is closed under conjunction (i.e. that if both it appears to *a* that *p* and it appears to *a* that *q*, then it appears to *a* that both *p* and *q*).

The argument is valid. Its soundness is questionable. The premiss that puts its soundness in question is [2]. As I pointed out, it is far from clear that [2] is true. If one avoids assuming [2] as a premiss, what remains is an invalid argument. Of course, every invalid argument can be transformed into a valid one by adding a 'tacit' premiss.

18. *The roles of imitation*

In the final part of the *Statesman*, the concept of imitation is applied to entities of three types: to present-day politicians, who are described as imitations of the statesman (cf. 303C3 < T8, where present-day politicians are said to be images – recall[61] that 'image' and 'imitation' are mere stylistic variants)[62]; to present-day regimes, which are described as imitations of the genuine constitution (cf. 293E3-6 = T1; 297C1-4 = T2); and to laws, which are described as imitations of statesmanship (cf. 300C5-7 = T7). Some of these applications of the concept of imitation are explicitly connected. In particular, the first application of the concept of imitation is explicitly connected to the second. For, the Visitor offers an argument (cf. 303B8-C5 < T8) to show that present-day politicians are imitations of the statesman because the regimes over which they rule are imitations of the genuine constitution, over which the statesman rules[63]. Thus, the first application of the concept of imitation is explained by appealing to the second.

It is tempting to assume that a similar connection obtains between the second application of the concept of imitation and the third, i.e. that the second application of the concept of imitation is explained by appealing to the third. In other words, it is tempting to assume that the application of the concept of imitation to present-day regimes is explained by appealing to its application to laws, namely to assume that present-day regimes are

[61] Cf. above, text to n. 41.

[62] Cf. also 301B1 and 301C3, where monarchic rulers are said to 'imitate' the genuine statesman and to be 'imitations' of him.

[63] I examined this argument earlier: cf. above, section to n. 41 and section to n. 60.

imitations of the genuine constitution because the laws on which they all rely (even though some of them take no account of these laws) are imitations of statesmanship, on which the genuine constitution relies. Although the text does not explicitly affirm this, there are some hints that the Visitor could be reasoning along such lines. For, on the two occasions when he asserts that present-day regimes are imitations of the genuine constitution (at 293E3-6 = T1 and 297C1-4 = T2), the Visitor adds that some of these regimes imitate the genuine constitution for the better and others for the worse, and he then goes on to specify that the difference between imitating for the better and for the worse has to do with whether the regimes respect the laws or take no account of them: this suggests that the fact that present-day regimes are imitations of the genuine constitution is intimately linked to their reliance on laws.

Earlier[64] I argued that the reason why the Visitor treats present-day regimes as imitations of the genuine constitution is that they appear to be constitutions without being constitutions, and I pointed out that their appearing to be constitutions is revealed by their being ordinarily called 'constitutions'. If the last paragraph's tempting assumption is correct, then a more thorough explanation of why the Visitor treats present-day regimes as imitations of the genuine constitution may be put forward: just as the reason why present-day politicians are imitations of the statesman is that present-day politicians bear a certain relation (i.e. ruling) to objects (i.e. present-day regimes) that are imitations of an object (i.e. the genuine constitution) to which the statesman bears that relation (for the statesman rules over the genuine constitution), so also the reason why present-day regimes are imitations of the genuine constitution is that present-day regimes bear a certain relation (i.e. reliance) to objects (i.e. laws) that are imitations of an object (i.e. statesmanship) to which the genuine constitution bears that relation (for the genuine constitution relies on statesmanship).

If these considerations are on the right track, then the applications of the concept of imitation in the *Statesman* are, so to speak, 'boxed' in one another: present-day politicians imitate the statesman because they rule over objects (present-day regimes) that imitate the genuine constitution (over which the statesman rules), and these objects imitate the genuine constitution because they rely on further objects (laws) that imitate statesmanship (on which the genuine constitution relies). The first application of the concept of imitation is then explained by appealing to the second,

[64] Cf. above, paragraph to n. 7.

which is in turn explained by appealing to the third. According to this picture, the application of the concept of imitation to laws is the most fundamental and ultimately explains its other two applications.

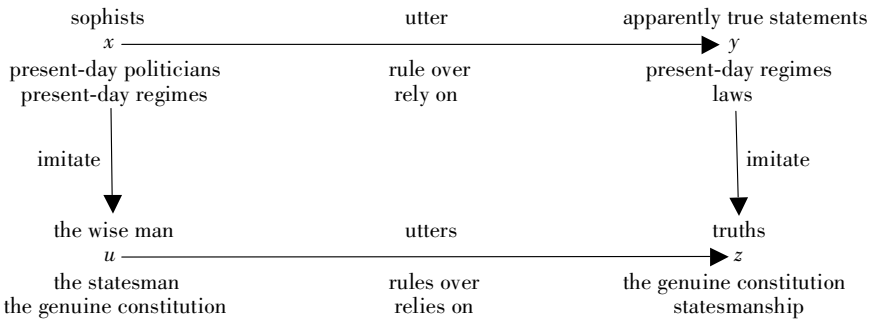
However, the tempting assumption about a connection between the second and the third application of the concept of imitation must remain a speculative suggestion. For, the evidence in its support is scarce. Moreover, at least one alternative possible account of the connection between the second and the third application of the concept of imitation should be mentioned: it cannot be excluded that while the present-day regimes that imitate the genuine constitution *for the better* imitate it because they rely on laws that in turn are imitations of statesmanship, on which the genuine constitution relies, the present-day regimes that imitate the genuine constitution *for the worse* imitate it because their ignorant rulers put themselves above the law like the rulers of the genuine constitution, namely genuine statesmen. In this case, Plato would be offering two different explanations of why present-day regimes imitate the genuine constitution: the imitation of present-day regimes that imitate for the better would be different with respect to the imitation of present-day regimes that imitate for the worse. The remarks of the Visitor at 301A10-B3 and at 301B10-C4 seem to go in this direction: in the first passage the Visitor speaks of the king who 'rules according to laws, imitating the one who has knowledge' (301A10-B1); in the second he speaks of the tyrant who 'acts neither according to laws nor according to customs' (301B10) but 'pretends to act like the one who has knowledge, saying that one must do what is best outside the written rules' (301C1-2). To be sure, these two passages speak of rulers who are imitating the genuine statesman, not of regimes that are imitating the genuine constitution; but what the two passages say about rulers imitating the genuine statesman easily translates into claims about regimes imitating the genuine constitution, and it suggests different explanations of what it is for a regime to imitate the genuine constitution for the better and what it is for it to perform such an imitation for the worse.

19. *The parallel structure of three arguments about imitations*

In the *Sophist*, the Visitor claims that sophists appear to be wise because they utter spoken imitations (statements) that can delude inexperienced youths into thinking that they are what they imitate, namely true statements. He also asserts (at 233C8) that sophists really are not wise,

and he describes them as imitators of the wise man (at 268c1). Given that in this theoretical context forms of 'to imitate φ' introduce an idea that could also be conveyed by the corresponding forms of 'to appear to be φ without being φ' or 'to instil the illusion of being φ'[65], the thesis of the Visitor may be plausibly taken to be that sophists imitate the wise man without being wise because they utter false but apparently true statements, namely statements that imitate true statements.

If the last section's tempting assumption about the connection between the second and the third application of the concept of imitation is correct, we obtain parallel explanations of three applications of the concept of imitation, the first to sophists, the second to present-day politicians, and the third to present-day regimes. The parallel explanations are illustrated by the following schema:



The relations pictured by three of the arrows in the schema (the *x*-*y* arrow, the *y*-*z* arrow, and the *u*-*z* arrow) explain the relation pictured by the remaining arrow (the *x*-*u* arrow). In general, *x*s imitate *u* because they bear a certain relation to *y*s, which imitate *z* or *z*s, to which *u* bears the same relation. By plugging in the expressions above the horizontal arrows, we obtain: sophists imitate the wise man because they utter apparently true statements, which imitate truths, which the wise man utters. By plugging in the expressions on the first line under the horizontal arrows, we obtain: present-day politicians imitate the statesman because they rule over present-day regimes, which imitate the genuine constitution, over which the statesman rules. By plugging in the expressions on the second line under the horizontal arrows, we obtain: present-day regimes imitate the genuine

[65] Cf. above, nn. 5 and 7 and the paragraphs to which they are appended.

constitution because they rely on laws, which imitate statesmanship, on which the genuine constitution relies.

Earlier[66] I questioned the soundness of the argument for the claim that present-day politicians imitate the statesman (i.e. appear to be statesmen but are not statesmen). Similar doubts may be raised about the two parallel arguments, the one for the claim that sophists imitate the wise man and the one for the claim that present-day regimes imitate the genuine constitution.

20. *Imitation in Plato's late philosophy*

The concept of imitation plays many roles throughout Plato's reflections. In the dialogues of the middle period (*Phaedo*, *Symposium*, *Republic*, *Phaedrus*) Plato puts it to work in order to explain the relation of participation of perceptible particulars to forms. However, in the late critical dialogues, this role is (to say the least) less prominent. This tendency is particularly clear in the *Sophist* and the *Statesman*, whose cosmological sections, despite their obvious echoes of the *Timaeus*, do not present the forms as paradigms of which perceptible particulars are imitations. The *Sophist* and the *Statesman* find other, more mundane but nevertheless important roles for the concept of imitation, which they employ to explain the nature of sophists, present-day politicians, present-day states, and laws. Thus, the concept of imitation remains a fundamental tool in Plato's philosophical machinery.

References

Adorno F. (trans. and comm.) (1988), *Platone, Dialoghi politici e lettere*, 3rd ed., Utet, Turin.

Annas J., Waterfield R. (trans. and comm.) (1995), *Plato, Statesman*, Cambridge University Press, Cambridge.

Anton J.P., Preus A. (eds.) (1989), *Essays in Ancient Greek Philosophy*, III, State University of New York Press, Albany.

Apelt O. (ed. and comm.) (1897), *Platonis Sophista*, Teubner, Leipzig.

Ast F. (1835-38), *Lexicon Platonicum sive Vocum Platonicarum Index*, Weidmann, Leipzig.

[66] Cf. above, section to n. 60.

Aubenque P., Narcy M. (eds.) (1991), *Études sur le* Sophiste *de Platon*, Bibliopolis, Naples.

Barnouw J. (2002), *Propositional Perception.* Phantasia*, Predication and Sign in Plato, Aristotle and the Stoics*, University Press of America, Lanham-New York-Oxford.

Bluck R.S. (1963), *Plato's Sophist: A Commentary*, ed. by G.C. Neal, Manchester University Press, Manchester 1975.

Bondeson W. (1972), *Plato's* Sophist*: Falsehood and Images*, in «Apeiron», 6, pp. 1-6.

Bossi B., Robinson T.M. (eds.) (2013), *Plato's* Sophist *Revisited*, De Gruyter, Berlin-Boston.

Brown L. (2010), *Definition and Division in Plato's* Sophist = Charles (2010), pp. 151-171.

Burnet J. (ed.) (1900-07), *Platonis Opera*, Clarendon Press, Oxford.

Campbell L. (ed. and comm.) (1867), *The Sophistes and Politicus of Plato*, Clarendon Press, Oxford.

Charles D. (ed.) (2010), *Definition in Greek Philosophy*, Oxford University Press, Oxford.

Cordero N.L. (trans. and comm.) (1993), *Platon, Le Sophiste*, Flammarion, Paris.

Cornford F.M. (1935), *Plato's Theory of Knowledge: The* Theaetetus *and the* Sophist *of Plato Translated with a Running Commentary*, Routledge and Kegan Paul, London-New York.

Destrée P., Herrmann F.-G. (eds.) (2011), *Plato and the Poets*, Brill, Leiden-Boston.

El Murr D. (2014), *Savoir et gouverner: Essai sur la science politique platonicienne*, Vrin, Paris.

Fischer J.F. (ed. and comm.) (1774), *Platonis Dialogi Tres: Sophista, Politicus, Parmenides*, Lngehem, Leipzig.

Fowler H.N., Lamb W.R.M. (trans.) (1925), *Plato, The Statesman, Philebus, Ion*, Heinemann and G.B. Putnam's Sons, London-New York.

Fraccaroli G. (trans. and comm.) (1911), *Platone, Il sofista e l'uomo politico*, 2nd ed., La Nuova Italia, Florence.

Gill C. (1995), *Rethinking Constitutionalism in* Statesman, pp. 291-303 = Rowe (1995b), pp. 292-305.

Gill M.L. (2010), *Division and Definition in Plato's* Sophist *and* Statesman = Charles (2010), pp. 172-199.

Giorgini G. (trans. and comm.) (2005), *Platone, Politico*, RCS, Milan.

Griswold C.L., Jr. (1989), *Politike Episteme in Plato's «Statesman»* = Anton and Preus (1989), pp. 141-167.

Hirsch U. (1995), *Μιμεῖσθαι und verwandte Ausdrücke in Platons* Politikos = Rowe (1995b), pp. 184-149.

Jowett B. (trans.) (1892), *The Dialogues of Plato*, 3rd ed., Oxford University Press and Humphrey Milford, Oxford-London.

Kamlah W. (1963), *Platons Selbstkritik im Sophistes*, Munich.

Lane M.S. (1995), *A New Angle on Utopia: the Political Theory of the* Statesman = Rowe (1995b), pp. 276-291.

Ledesma F. (2009), *Le logos du* Sophiste*. Image et parole dans le* Sophiste *de Platon*, in «Elenchos», 30, pp. 207-253.

Long A.G. (2013), *Conversation and Self-Sufficiency in Plato*, Oxford University Press, Oxford.

LSJ = Liddell H.G., Scott R., Jones H.S. (1996), *A Greek-English Lexicon*, 9th ed., with a Revised Supplement, Clarendon Press, Oxford.

Márquez X. (2012), *A Stranger's Knowledge. Statesmanship, Philosophy, and Law in Plato's* Statesman, Parmenides Publishing, Las Vegas-Zurich-Athens.

Marušič J. (2011), *Poets and Mimesis in the* Republic = Destrée and Herrmann (2011), pp. 217-240.

Napolitano Valditara L.M. (2007), *Platone e le 'ragioni' dell'immagine. Percorsi filosofici e deviazioni tra metafore e miti*, Vita e Pensiero, Milan.

Notomi N. (1999), *The Unity of Plato's* Sophist*: Between the Sophist and the Philosopher*, Cambridge University Press, Cambridge.

Notomi N. (2011), *Image-Making in* Republic *X and the* Sophist*: Plato's Criticism of the Poet and the Sophist* = Destrée and Herrmann (2011), pp. 299-326.

Palumbo L. (1995), *Realtà e apparenza nel* Sofista *e nel* Politico = Rowe (1995b), pp. 175-183.

Palumbo L. (2013), Mimesis *in the* Sophist = Bossi and Robinson (2013), pp. 269-278.

Pellegrin P. (1991), *Le* Sophiste *ou de la division. Aristote-Platon-Aristote* = Aubenque and Narcy (1991), pp. 389-416.

Pippin R.B. (1979), *Negation and Not-Being in Wittgenstein's* Tractatus *and Plato's* Sophist, in «Kant-Studien», 70, pp. 179-196.

Pradeau J.-F. (2009), *Platon, l'imitation de la philosophie*, Aubier, Paris.

Rickless S.C. (2010), *Plato's Definition(s) of Sophistry*, in «Ancient Philosophy», 30, pp. 289-298.

Robinson D.B. (1995), *The New* Oxford Text *of Plato's* Statesman*: Editor's Comments* = Rowe (1995b), pp. 37-46.

Rowe C.J. (ed., trans., and comm.) (1995a), *Plato, Statesman*, Aris & Phillips, Oxford.

Rowe C.J. (ed.) (1995b), *Reading the* Statesman*. Proceedings of the III Symposium Platonicum*, Academia Verlag, Sankt Augustin.

Rowe C.J. (1995c), *Introduction* = Rowe (1995b), pp. 11-28.

Rowe C.J. (1997), rev. of Annas and Waterfield (1995), in «Classical Review», n.s. 47, pp. 277-279.

Rowe C.J. (trans. and comm.) (1999), *Plato, Statesman*, Hackett, Indianapolis-Cambridge (MA).

Rowe C.J. (2001), *Killing Socrates: Plato's Later Thoughts on Democracy*, in «Journal of Hellenic Studies», 121, pp. 63-76.

Rowe C.J. (2005), *The* Politicus *and Other Dialogues* = Rowe et al. (2005), pp. 233-257.

Rowe C.J., Schofield M., Harrison S., Lane M. (eds.) (2005), *The Cambridge History of Greek and Roman Political Thought*, Cambridge University Press, Cambridge.

Ryle G. (1966), *Plato's Progress*, Cambridge University Press, Cambridge.

Sabine G.H. (1962), *A History of Political Theory*, 3rd ed., Holt, Rinehart and Winston, New York.

Sallis J. (ed.) (2017), *Plato's* Statesman*: Dialectic, Myth, and Politics*, State University of New York Press, Albany.

Skemp J.B. (trans. and comm.) (1952), *Plato's Statesman*, Routledge and Kegan Paul, London.

Stallbaum J.G. (ed. and comm.) (1841), *Platonis Politicus et Incerti Auctoris Minos*, Hennings and Black and Armstrong, Gotha-London.

Stephanus H. (ed.) (1578), *Platonis Opera Quae Extant Omnia*, Henricus Stephanus, Geneva.

Taylor A.E. (trans.) (1961), *Plato, The Sophist and the Statesman*, ed. by R. Klibansky and E. Anscombe, Thomas Nelson, London.

Teisserenc F. (2005), *«Il ne faut en rien être plus savant que les lois». Loi et connaissance dans le* Politique, in «Les Études philosophiques», pp. 367-383.

Vernant J.-P. (1975), *Naissance d'images* = Vernant (1979), pp. 105-137.

Vernant J.-P. (1979), *Religions, histoires, raisons*, Maspero, Paris.

Warrington J. (trans.) (1961), *Plato, Parmenides, Theaitetos, The Sophist, The Statesman*, Dent and Dutton, London-New York.

Wolff F. (1991), *Le chasseur chassé. Les définitions du sophiste* = Aubenque and Narcy (1991), pp. 17-52.

Zadro A. (1961), *Ricerche sul linguaggio e sulla logica del* Sofista, Antenore, Padua.

Abstract

In the Statesman *Plato identifies the art of the statesman with a highly specialized branch of knowledge. It is this knowledge that must have the highest authority in the state: all other forms of organized society are merely imitations of the society based on the statesman's knowledge, which is the only genuine constitution. The concept of imitation is applied not only to describe the relationship between the genuine constitution and other types of organized society, but also to the relationship between the statesman and everyday politicians and to the relationship between the statesman's knowledge and law. It turns out that these three applications of the concept of imitation are reciprocally connected. Plato explicitly argues that everday politicians are imitations of the genuine statesman* because *everyday societies are imitations of the genuine constitution. This study explores the possibility that everyday societies could be imitations of the genuine constitution* because *law is an imitation of the statesman's knowledge.*

Keywords: knowledge; statesmanship; politician; constitution; imitation; law; sophistry.

Paolo Crivelli
Département de philosophie - Université de Genève
paolo.crivelli@unige.ch

T

Che cos'è un atto d'impegno? Husserl e Reinach sul "soggetto di livello superiore" (Noi) e gli atti (non) sociali[1]

Petar Bojanić

Introduzione

Sugli atti sociali, oggetto precipuo dell'epistemologia sociale e dell'ontologia sociale, non si è scritto moltissimo. Edmund Husserl e il suo allievo Adolf Reinach sono stati tra i primi sia a parlarne esplicitamente, sia a teorizzare ciò ch'essi definiscono "atto sociale negativo". Lo stesso Husserl fu probabilmente il primo a dilungarsi sul problema posto dalla prima persona plurale "Noi"[2], ossia dai "gradi superiori della comunità intermonadica"[3], un tema che è diventato nel tempo il centro dell'ontologia sociale.

Il mio scopo è introdurre e descrivere un tipo di atto – ossia dimostrare che la tematizzazione di tale atto è giustificata dal suo essere un atto sociale – che vorrei definire "atto d'impegno". Non è certo se l'impegno sia parte di ciò che Husserl e Reinach chiamano "atto sociale" – nel senso di un passaggio che un atto sociale può implicare e comprendere, ma non necessariamente – o piuttosto un caso del tutto particolare di atto sociale, la cui funzione sarebbe quella di istituzionalizzare un gruppo, ossia di convertirlo in istituzione.

La mia proposta consiste in un breve rimando a una scena *ur*-istituzionale, una delle più importanti fantasie visive dell'Occidente: la costruzione della torre di Babele. Il capitolo 11 del primo libro della Torah (*Bereshit*)

[1] Traduzione di Ernesto C. Sferrazza Papa.

[2] Per una problematizzazione del "Noi" in Husserl si veda su tutti il classico testo di M. Theunissen, *Der Andere: Studien zur Sozialontologie der Gegenwart*, De Gruyter, Berlin 1965.

[3] Cfr. E. Husserl, *Meditazioni cartesiane* (1950), a cura di F. Costa, Bompiani, Milano 2009, § 56, p. 147.

 DOI: 10.4454/teoria.v39i1.64

narra di un gruppo di migranti arrivato in un luogo nuovo. Questo gruppo parla un unico linguaggio, lavora e costruisce insieme[4]. I suoi membri parlando fra di loro si mobilitano costantemente ("venite", "facciamo un mattone", "costruiamoci una città", "diamoci un nome"). Il gruppo ha un progetto comune e, a un certo punto, decide di istituzionalizzarsi per non frantumarsi. Il gruppo ha un solo obiettivo e in due mosse conduce almeno sei distinte operazioni (l'autore del capitolo le divide per poi metterle insieme apparentemente alla rinfusa, sebbene la natura della narrazione e la loro enumerazione sembra renderle temporalmente sequenziali).

Le prime tre sono simultanee: 1) il parlare comprendendosi reciprocamente ("si parlano l'uno con l'altro"); 2) il mutuo incoraggiamento attraverso il linguaggio, ottenuto mediante l'uso dell'imperativo ("venite"; *havah*), nonché la mobilitazione e la disponibilità degli attori a mantenere l'(auto)consapevolezza di essere parti di un più ampio tutto; (3) l'invenzione di un nuovo tipo di edificio.

Le seconde tre azioni simultanee sono: 1) il concentrarsi, il preoccuparsi e il muoversi insieme, congiuntamente ("facciamo"); 2) l'intenzionalità collettiva e la creazione di un grandioso schema, del progetto di un edificio comune; 3) la creazione di un'istituzione (documentazione, società) nominata e in tal modo riconosciuta come un'entità indipendente.

Ovviamente, è chiarissimo che questo gruppo (il suo attributo è l'uso del pronome "noi", che corrisponde all'imperativo: "facciamo") comprende individui di genere ed età differenti (una pluralità di Io), nessuno dei quali risulta in qualche modo distinto. Questa scena teatrale (della quale tutti noi conosciamo la fine, ossia la figura/persona che sale sul palco e nega il permesso di costruzione per l'impresa) potrebbe aiutarci sia a distinguere con maggiore precisione un gruppo da un'istituzione (il "Noi" del gruppo e quello di questa nuova entità di livello superiore), sia a distinguere meglio l'empatia dagli atti sociali prodotti dai soggetti di Husserl. Solo ed esclusivamente gli atti sociali o alcuni atti sociali molto speciali, simultanei e reciproci[5], potrebbero costruire ciò che gli antichi

[4] Questo gruppo pre-dato, o questo pre-dato "Noi", ritorna in un passaggio della *Crisi*: «nella vita che conduciamo insieme noi abbiamo in comune un mondo già dato, il mondo che è e che vale per noi, il mondo di cui noi, anche nel nostro vivere-insieme, facciamo parte, il mondo per tutti noi, il mondo già dato in questo senso d'essere» (E. Husserl, *La crisi delle scienze europee e la fenomenologia trascendentale* (1959), a cura di E. Filippini, il Saggiatore, Milano 1962, p. 139).

[5] «Io li ottengo piuttosto nel senso di una comunità umana, dell'*uomo* stesso il quale già come individuo ha il senso di un membro della comunità (il che si estende alle società animali); ora è proprio di questo senso di comunità umana il rapporto costituito dall'essere l'uno per l'altro,

costruttori designano con la parola "nome", ciò che noi oggi chiamiamo "istituzione", e che inizialmente Husserl chiama "il collettivo"[6].

La mia intenzione è abbozzare la descrizione di un tipo di "atto sociale" che ho chiamato "atto impegnato" (e che dovrebbe essere differente dall'"impegno congiunto"). Vorrei disvelare e demarcare questi atti impegnati all'interno del tentativo di Husserl di definire e istituire, *de facto*, atti sociali di questo tipo. I corollari di quest'operazione sarebbero: mostrare l'importanza degli atti sociali nella costruzione di un qualche tipo di nuova entità, che è sempre problematica da nominare (una delle quali è certamente "Noi"); distinguere il più chiaramente possibile gli atti sociali dall'empatia; definire alcuni atti sociali "atti impegnati" in modo da alleggerire e chiarire gli sforzi di Husserl nella determinazione degli atti sociali; rivalutare il contributo di Adolf Reinach nella definizione degli atti sociali in confronto a Husserl. Ritengo quest'ultimo punto particolarmente interessante perché, mentre permette di focalizzarsi meglio sul significato della reciprocità come caratteristica fondamentale degli atti sociali per Husserl, riscopre l'importanza e l'originalità di Reinach.

1. *L'atto sociale negativo in Reinach e Husserl*

Nel breve schizzo *Nichsoziale und soziale Akte*[7], Reinach offre una brillante definizione degli atti sociali, respingendo ed escludendo tutto ciò

rapporto che determina la parificazione del mio esserci con quello di ogni altro» (E. Husserl, *Meditazioni cartesiane*, cit., p. 148).

6 «Dobbiamo allora considera la seguente questione: il matrimonio, l'amicizia, queste sono unità collettive "nate" fuori dalla relazione "psicologica" della pluralità di persone e che quindi le connette in un'unità superiore» (E. Husserl, *Zur Phänomenologie der Intersubjektivität. Texte aus dem Nachlass. Erster Teil: 1905-1920*, hrsg. von I. Kern, Martinus Nijhoff, Leiden 1973, p. 101). Husserl usa il termine "collettivo" nel 1910 ed espande il significato di questo termine per includervi la famiglia, le gilde professionali, le associazioni, le corporazioni, le religioni e in generale ogni tipo di istituzione (cfr. *ivi*, p. 98). Successivamente, Husserl sostituisce il termine "collettivo" con *neue Objektivitäten höherer Stufe*, privo di accenni alla naturalità (cfr. Id., *Zur Phänomenologie der Intersubjektivität. Texte aus dem Nachlass. Zweiter Teil: 1921-1928*, hrsg. von I. Kern, Martinus Nijhoff, Leiden 1973, p. 192).

7 Cfr. A. Reinach, *Nichtsoziale und soziale Akte* (1911), in Id., *Sämtliche Werke*. Kritische Ausgabe und Kommentar, hrsg. von K. Schumann und B. Smith, Philosophia Verlag, München 1989, pp. 355-360). Reinach non spiega mai esplicitamente cosa sia un *Nichtsoziale Akte* (negli appunti sono menzionati due o tre volte come opposti agli atti sociali), e nemmeno lo fanno i suoi commentatori. Tuttavia, egli tratta gli atti sociali anche nel libro del 1913 *Die apriorischen Grundlagen des bürgerlichen Rechtes*, dove sostanzialmente parafrasa le osservazioni del 1911.

che non rientra nella definizione chiamandolo *nichtsoziale Akte* – che di conseguenza rappresenta un refuso irrilevante, nel senso che tutto ciò che è un atto non sociale non è un atto sociale: non dovrebbe dunque esserci una nuova speciale entità chiamata *nichtsoziale Akte*[8]. Tuttavia, proprio a ridosso di questa definizione, Reinach complica la questione chiamando in causa tale distinzione attraverso l'esempio della preghiera:

> La forma apparente esiste solo perché le cose sono tali che noi possiamo conoscere solo i nostri atti interni attraverso le loro forme apparenti. La preghiera, per esempio, *è* un atto sociale. Lì, il primo esiste, mentre il secondo no (non ha una forma fenomenica). L'uomo religioso assume che il destinatario ascolti la preghiera senza la forma fenomenica. Di conseguenza, è possibile anche una preghiera silenziosa[9].

La preghiera *è* (l'*ist* nell'originale è in corsivo) un atto sociale perché, a dispetto dell'incertezza nell'esistenza di colui che ascolta la preghiera (qui seguo il realismo di Reinach) e con il quale non c'è realmente una connessione, nondimeno c'è un atto interno o un'esperienza, così come il destinatario di colui (l'uomo religioso) che prega assume non solo l'esistenza, ma accetta anche ciò che viene inviato (anche se il credente non manda alcunché, ad esempio una serie di parole inespresse prese da un protocollo familiare). Reinach continua: «*jeder soziale Akt gründet in einem Erlebnis, das nicht sozialer Akt ist*»; la negazione così costruita indica che dietro a ogni atto sociale c'è realmente un qualche tipo di esperienza che non è un atto sociale, o che ogni atto sociale è necessariamente fondato in un atto non sociale, o ancora che l'esperienza in quanto tale (presa isolatamente) è un atto non sociale.

In aggiunta all'origine dell'atto sociale individuata in qualcosa che non è un atto sociale (o che è un atto non sociale), l'atto sociale non solo non deve, in questo caso, avere una forma apparente (di contro, per Husserl o

[8] In un paio passaggi, del tutto in sintonia con il capitolo di Reinach sugli atti sociali, possiamo classificare la loro differenza dagli atti sociali negativi: un atto sociale è spontaneo (uno negativo no). La spontaneità designa le azioni interne del soggetto. Un atto sociale non è in pace in quanto tale, in se stesso: esso richiede di essere esternalizzato (uno negativo no). Un atto sociale penetra in un altro, mentre il negativo non lo fa. Un atto sociale può essere percepito, uno negativo no. Un atto sociale ha un aspetto interno ed esterno (fenomenico); un atto negativo, solamente interno. L'espressione di un atto sociale non è accidentale o involontaria, mentre quella del negativo sì. Un atto sociale può avere molti destinatari e destinatarie, quello negativo non può. Cfr. A Reinach, *Die apriorischen Grundlagen des bürgerlichen Rechtes*, Niemeyer, Halle 1913, pp. 158-161 e p. 164.

[9] Id., *Nichtsoziale und soziale Akte*, cit., p. 357.

Reinach solo atti non sociali non hanno una forma apparente), ma, in ultima analisi, non vi deve essere alcuna forma di risposta dal destinatario che ha ricevuto il messaggio (giacché questo sarebbe un altro atto non sociale, o la prova che l'altro, il destinatario, potrebbe essere autistico). Questa è per me il problema che l'esempio di Reinach della preghiera ci mostra. Sembra infatti che in quel caso non vi sia alcun bisogno di prove che il destinatario e la persona religiosa abbiano avuto una qualunque forma di relazione sociale. Un atto sociale deformato in tal modo, circondato da tutti i lati da differenti forme di negazione, è basato esclusivamente sull'assunzione che il messaggio è arrivato lì dove era destinato. L'uomo religioso può solo potenzialmente testimoniare di aver mandato un messaggio di preghiera, che è stato ricevuto e al quale potenzialmente si è ottenuta, allo stesso modo – ossia silenziosamente –, una risposta.

Tuttavia, la testimonianza (ossia la narratività, dal momento che Reinach distingue tra affermazione [*Behauptung*] e messaggio [*Mitteilung*])[10] dell'uomo religioso a proposito di questa azione indirizzata a un altro uomo rappresenterebbe un atto sociale di bassissimo valore, dal momento che non ha bisogno necessariamente di obbligare l'ascoltatore ad accordare le sue future azioni con ciò che ha ascoltato dall'uomo religioso. Nel suo libro, Reinach rielabora e rafforza questa scena di "preghiera":

> Immaginiamo una comunità che comprenda esseri in grado di percepire direttamente e immediatamente le loro mutue esperienze. Saremmo costretti ad ammettere che in una simile comunità gli atti sociali dovrebbero apparire qualcosa che possiede solo anima, ma non corpo. Se assumiamo che un essere al quale ci rivolgiamo nei nostri atti sociali sia in grado di afferrare immediatamente la nostra esperienza, noi umani dovremmo in quel modo rinunciare senza se e senza ma all'aspetto esteriore dei nostri atti sociali. Ricordate quella silenziosa preghiera indirizzata a Dio e cercate di manifestarla a lui. Ciò dovrebbe essere visto come un atto sociale puramente spirituale[11].

Gli atti sociali negativi hanno dunque una funzione regolativa. Da questo punto di vista, in una comunità ideale gli atti sociali negativi sono divenuti palesemente sociali, o sociali nel vero senso del termine. Più semplicemente: gli atti sociali negativi cesserebbero in questo caso di essere negativi. Ma cosa vi è di negativo negli atti sociali (nelle fondazioni degli atti sociali), o che cosa sono degli atti o delle azioni negative (se essi *sono*

[10] Id., *Die apriorischen Grundlagen des bürgerlichen Rechtes*, cit., p. 161.
[11] *Ibidem*.

atti, come ammoniva Gilbert Ryle[12])? La negazione, paradossalmente, presuppone socialmente quello che rifiuta. È importante allora descrivere meglio la trasformazione di Reinach dalla comprensione fenomenologica della preghiera come atto sociale nelle sue lezioni (*Nichtsoziale und soziale Akte* è in effetti una serie di note compilate da due uditori delle lezioni di Reinach), alla compilazione "giuridica" successiva.

Per il fenomenologo, al fine di essere un atto sociale in quanto tale, è sufficiente che vi sia un'esperienza, ossia un atto interiore, e lo stesso vale per il processo di destinazione all'altro, che non ha riguardo del fatto se il destinatario esista o meno. Reinach chiama in questione l'atto sociale se, ad esempio, l'esperienza è falsa o finta (nel caso dell'ipocrisa), il che potrebbe addirittura configurarsi come un caso peggiore, perché l'autenticità dell'esperienza non deve necessariamente influenzare l'efficienza o la performatività dell'atto o la sua risposta[13]. In Reinach ciò è simile alla tematizzazione della promessa. La produzione di obbligazione è qualche volta più importante per la costituzione del soggetto che promette rispetto all'insistenza di Reinach sull'eco nell'altro[14], vale a dire: sul significato dello scambio di promesse, e non solo, per la costituzione della comunità. Il giurista Reinach, dall'altro lato, riattualizza, anche se in maniera insufficiente, l'altro che non è indifferente alla mia promessa (o per esempio alle mie scuse) o impercettibile quando ascolta la mia preghiera.

Il passaggio «immaginiamo una comunità che comprenda esseri in grado di percepire direttamente e immediatamente le loro mutue esperienze», forse uno dei luoghi più cruciali del capitolo di Reinach sugli atti sociali, richiede esempi più complessi di quello della persona religiosa che si rivolge a un destinatario invisibile attraverso la preghiera. E non solo esempi ma, analogicamente, attività collettive di impegno che a tutti gli effetti trasformerebbero il non sociale in atti sociali, o altrimenti li eliminerebbero

[12] Così Ryle scrive: «quel che è interessante è la classe di atti (se *sono* atti) consistenti nella *non*-esecuzione intenzionale di determinate azioni. Ad esempio, *rimando* la scrittura di una lettera qualora, senza aver dimenticato tale compito, non scrivo ora la lettera pur potendo farlo; è in questo senso che io non posso rimandare la *tua* azione di scrivere una lettera» (G. Ryle, *«Azioni» negative»* (1973), in Id., *Pensare pensieri*, a cura di G. Melilli Ramoino, Armando, Roma 1990, p. 130).

[13] Cfr. A. Reinach, *Die apriorischen Grundlagen des bürgerlichen Rechtes*, cit., pp. 162-163. Cfr. F. de Vecchi, *The Existential Quality Issue in Social Ontology: Eidetics and Modifications of Essential Connections*, in «Humana. Mente Journal of Political Studies», 31 (2016), pp. 194-197.

[14] E. Husserl, *Zur Phänomenologie*, I, cit., p. 99.

del tutto[15]. Se immaginiamo una comunità in cui gli atti sociali negativi sono eliminati nell'interazione reciproca[16], allora ciò chiamerebbe in questione una delle più importanti distinzioni tra atto sociale e atto non sociale fornite da Reinach: un atto sociale può avere molti destinatari e destinatarie mentre, dall'altro lato, un atto negativo no.

Due esempi (o piuttosto: un esempio che assume due forme) potrebbero correggere il passaggio citato da Reinach. Il primo è l'evocazione dei fantasmi (l'esempio è di Reinach). In questo caso ciò che il gruppo compie è precisamente uno sforzo comune mediante l'applicazione di atti sociali individuali negativi nello sforzo d'inviare un messaggio che manifesterà esseri nient'affatto sociali o socievoli. Il secondo è la preghiera collettiva di "individui religiosi" (che potrebbero anche essere silenziosi), che nel loro silenzioso messaggio vivono e costruiscono una nuova comunità di tutti.

In contrasto con il suo allievo, Husserl non riconosce e non accetta qualsivoglia forma di asimmetria o di non reciprocità tra attori sociali. Anche quando si eserciterà, e lo farà molte volte negli appunti presi nella fase matura della sua vita, con l'idea che l'empatia possa anche essere realmente reciproca e attiva, Husserl rifiuterà una simile concezione perché io non posso vedere se gli altri simultaneamente mi notano, o se stiano osservando loro stessi, o ancora se sono tutti quanti interessati a me quando mi rivolgo a loro. Affinché abbia luogo un'autentica *communicatio*[17], la mia attività necessita di essere esplicitamente dichiarata e ricambiata (ossia: attivamente orientata verso di me)[18]. Husserl insiste sulle parole attività (*Aktivität*), immediatezza e progetto (*Vorhabe*), nonché sulla volontà che qualcosa venga esplicitamente dichiarata all'altro (o agli altri) quali condizioni principali affinché l'unione comunicativa e sociale abbia luogo.

Quali sono allora le caratteristiche principali degli atti sociali, e fino a

[15] Una delle principali caratteristiche degli atti d'impegno è l'eliminazione del negativo, del non impegnato, del non sociale, dell'antisociale.

[16] Cfr. E. Husserl, *Zur Phänomenologie*, II, cit., pp. 192-205. Più tardi Husserl parlerà di varie forme di espressione linguistica, dell'aspettarsi che colui al quale rivolgiamo il messaggio risponda, e del disappunto quando la risposta va perduta. Al fine di ricevere una risposta, è necessario offrire il silenzio a uno di coloro dai quali ci attendiamo una risposta, che in un certo senso equivale a un atto negativo (cfr. Id., *Zur Phänomenologie der Intersubjektivität. Texte aus dem Nachlass. Dritter Teil: 1929-1935*, hrsg. von I. Kern, Martinus Nijhoff, Leiden 1973, pp. 474-475).

[17] Id., *Zur Phänomenologie*, II, cit., p. 199.

[18] Id., *Zur Phänomenologie*, III, cit., p. 472. Sul rapporto tra empatia e atti sociali, sulle reciproche o mutue relazioni sociali, così come sulla reciproca o unilaterale empatia, si veda T. Szanto, *Husserl on Collective Intentionality*, in A. Salice, H.B. Schmid (eds.), *The Phenomenological Approach to Social Reality: History, Concepts, Problems*, Springer, Cham 2016, pp. 4-5.

che punto è possibile fare uso delle molteplici designazioni husserliane al fine di tracciare una nuova e più modesta distinzione tra, per esempio, atti sociali e istituzionali, o tra atti sociali e atti d'impegno? Per dirla in altri termine: l'appena abbozzato tentativo di Husserl di fondare la comunità negli atti sociali amplia eccessivamente il loro significato, e dunque lo indica anche imprecisamente?

In una prima fase (molto precoce, intorno al 1910) Husserl insiste sul fatto che le relazioni intersoggettive sono in loro stesse reali e che gli individui che le conducono sono reali[19]. Per di più, gli atti, gli atti comunicativi (*kommunikative Akte*), gli "atti indirizzati ad altri" (*die sich an den Anderen wenden*) implicano che l'altro sia a conoscenza del venire interpellato. L'altro ha bisogno di comprendere il mittente da cui certi atti sono stati inviati e rispondere con un atto dello stesso tipo. Così Husserl: «questi sono atti che producono una più alta unità di consapevolezza da persona a persona»[20].

L'altra caratteristica degli atti sociali su cui mi vorrei soffermare, e che Husserl ingegnosamente sviluppa, si riferisce alla norma della comunità (*eine Gemeinschaftsnorm*), o norma comune. Husserl ritiene che gli atti sociali non siano davvero norme che obbligano, ma piuttosto pseudo-norme o pseudo-obbligazioni[21]:

> è una volontà di regolazione comune, riconosciuta dagli individui ed è superindividuale. […] Colui che non risponde a un saluto, colui che non ringrazia è un villano. […] Se io mi rivolgo a qualcuno cortesemente, ho il diritto di aspettarmi una cortese risposta da parte sua, che sia magari un ringraziamento come risposta a un saluto cordiale, e così via[22].

Nel 1921 Husserl tenta un'altra strada: nega gli atti sociali che non sono atti sociali o che non sono ancora atti sociali. Innanzitutto, l'amore: il mio amore o la mia ammirazione non è ancora un atto sociale di un qualche tipo. Se io amo, non c'è ancora un atto di amore sociale (*Akte der*

[19] E. Husserl, *Zur Phänomenologie*, I, cit., pp. 96-97.

[20] *Ivi*, p. 98.

[21] Altra caratteristica decisiva degli atti d'impegno: non vincolare mai, bensì mantenere la capacità di "dare inizio", oppure quella di circolare lungo le attività che connettono con l'altro o con gli altri. Nella XII Lezione del suo libro più importante, John Austin descrive protocolli ("l'impegno" e il "pegno" sono alcuni di essi) che non sono vincolati – come la "promessa" –, e tuttavia possono comunque «impegnarti a fare qualcosa, ma includere anche dichiarazioni o gli annunci riguardo alle proprie intenzioni» (J. Austin, *Come fare cose con le parole. Le «William James Lectures» tenute alla Harvard University nel 1955*, a cura di C. Penco e M. Sbisà, Marietti, Genova 1987, p. 110).

[22] E. Husserl, *Zur Phänomenologie*, I, cit., pp. 105-106.

sozialen Liebe). Se io faccio deliberatamente qualcosa per un'altra persona al fine di fare notare il mio comportamento, il modo in cui sto esponendo me stesso, nessuno di questi è ancora un atto sociale. È un atto sociale fare qualcosa sperando che l'altro, notando la mia intenzione, risponderà a suo modo[23]. Nello stesso anno, Husserl scopre qualcosa di nuovo analizzando la famiglia. Egli dimostra come un'unità temporanea (*vorübergehende Gemeinschaft*) si trasformi in un'istituzione ordinata (*geregelten Institution*) se i suoi membri abitudinariamente mangiano insieme. I pasti insieme (atti sociali) sono gli elementi base per l'istituzione della famiglia come istituzione sociale. Per una famiglia, al fine di essere una famiglia (per un "Noi" diventare un'istituzione), è insufficiente vivere fianco a fianco:

> molto più di quello, è una questione del modo in cui si vive insieme, del modo di essere in relazione con l'altro, mutualmente reciproci all'interno di situazioni di vita, agendo in maniera da influire reciprocamente gli uni sugli altri, in relazioni che funzionano reciprocamente, e basate su ciò che dell'azione dell'uno penetra nell'azione dell'altro[24].

L'ultima caratteristica degli atti sociali, per come l'ha immaginata e definita Husserl, molto probabilmente la sua più grande scoperta, appare all'improvviso nel gennaio 1931, quando Husserl nomina questo altro come "il terzo" (*dritte*). Così formulato, il "terzo" diviene la premessa per la scoperta del "Noi", ossia dell'atto istituzionale: «il mio vicino (qualcuno di vicino a me), che adesso percepisco è già adesso il terzo […] che aiuta nella formazione continua del mondo dallo stato-del-mondo iniziale "per noi due" a un mondo per noi tre»[25]. Il momento finale che per Husserl determina l'atto sociale apre la questione sulla natura della reciprocità. Se vi sono due persone "in relazione reciproca l'una con l'altra", e se abbiamo precedentemente definito questa situazione come il loro impegno, allora il loro sviluppare e incoraggiare l'interazione implica l'apparire del terzo o del gruppo, e dunque dell'istituzione?

[23] Id., *Zur Phänomenologie*, II, cit., p. 166. Affinché un atto sociale sia tale è necessario un processo a catena non mimetico. Ogni persona successiva, ogni persona che prosegue un atto di qualcun altro conferma che l'atto è sociale. Se il mio atto impegna un altro il cui atto a quel punto impegna me o qualche terzo, allora il mio atto è sociale. La socialità di un atto è decisa dunque da ognuno degli atti che lo seguono. Cfr. *ivi*, p. 193.

[24] *Ivi*, p. 179.

[25] E. Husserl, *Zur Phänomenologie*, III, cit., p. 134.

2. *Atti impegnati*

Chiamo questi atti "impegnati", soprattutto perché essi cambiano l'istituzione (l'alterano ma simultaneamente la creano) introducendo nuove regole. Anche se talvolta è estremamente difficile sviluppare o differenziare un'azione o un evento, un atto impegnato è quello che crea qualcosa di nuovo (il "terzo" nel vocabolario di Husserl), qualcosa come un evento reale. Tali atti producono una certa forma di obbligazione in tutti i membri di un gruppo (e in quelli che non lo sono ancora diventati; vale a dire, è imperativo impegnare tutti), ossia obbligano il gruppo in quanto tale (l'agente-gruppo) a formare nuovi tipi di obbligazione. Essere impegnato significa fare affidamento su tutti gli altri e lavorare in modo tale da produrre un'ampia partecipazione, un onere (un pegno, un impegno) che dovrebbe reiterare l'obbligazione e la responsabilità istituzionale anche in quelli che sono stati marcati come sbandati e sovversivi, coloro che commettono atti negativi o sovversivi.

Di conseguenza, insisto sul fatto che vi sia un numero impreciso e incerto di differenti e inclassificate attività che hanno la capacità di:

a) non solo incoraggiare o spingere o attivare l'altro (o gli altri) verso identiche o simili azioni o reciproche reazioni, ma anche produrre una pseudo-obbligazione che implica un'azione congiunta di gruppo ("fare qualcosa come gruppo");
b) non solo obbligare membri del gruppo a compiere qualcosa insieme, ma financo a eccedere i confini dell'impegno comune del gruppo, obbligando *a priori* i non membri o tutti i potenziali e futuri partecipanti attraverso un'azione impegnata e coordinata.

Come si costituiscono dunque tali azioni, quelle che coinvolgono gli altri (tutti gli altri) o che hanno la capacità di impegnare (di tenere insieme, di raccogliere e legare anche quelli che non sono simultaneamente presenti in un unico luogo)? Descriviamo, elenchiamo, assumiamo un paio di significati dei verbi "impegnare" e "coinvolgere". Questi tre verbi alla prima persona plurale dell'imperativo ("descriviamo", "elenchiamo", "assumiamo"), che potrebbero essere pronunciati a gran voce da ogni singolo individuo nello stesso momento sospendendo il proprio parlare in prima persona singolare (solo "Noi" può sostituire "Io"; e solo "Io" può pronunciare il pronome "Noi"), potrebbero rappresentare insieme un tipo di obbligazione per coloro che sono potenzialmente a portata d'orecchio e comprendono l'enunciazione. Il modo in cui questi verbi sono usati potenzialmente

connette, mobilita e invita gli altri all'accordo o all'azione individuale, e simultaneamente (anche) li invoca alla (stessa/comune) risposta. La loro risposta comune o la loro azione comune è confermata non solo quando ognuno conduce una certa attività data (ad esempio descrivere, assumere o elencare i significati delle parole "coinvolgere" e "impegnare") oppure quando simultaneamente e con coinvolgimento, abbandono e attività di concerto rappresenta la performance collettiva di "assumere", "descrivere" ed "elencare". È anche confermata quando questi tre imperativi sono ripetuti o semplicemente pronunciati: "descriviamo e assumiamo ed elenchiamo". La prima persona plurale è una delle iniziali, ma incondizionate, condizioni di istituzionalizzazioni del lavoro di gruppo o di impegno comune. Certamente non la sola. Verbi come domandare, suggerire, implorare, supplicare, richiedere, esigere, ordinare, così come provare, argomentare, giustificare o difendere (non necessariamente usati all'imperativo) potrebbero incoraggiare all'impegno e potenzialmente all'impegno comune.

Un'azione impegnata sarebbe allora innanzitutto pubblica o proclamata (giacché non può essere un atto sociale negativo o un segreto, un'azione nascosta eseguita in silenzio). Inoltre, essa è per natura pro-vocatrice: è una chiamata o un messaggio a tutti gli altri, un suggerimento a tutti ad avvicinarsi, a unirsi (non solo ai membri di un gruppo, ma anche a quelli estranei e fuori dal gruppo), perché "impegnare" significa precisamente agire incoraggiando altri a fare qualcosa insieme, rendendoli così modo membri di un impegno futuro.

Tuttavia, la specificità dell'azione impegnata sta nel supporre questa forma di grandioso lavoro, di adesione ("sacrificarsi per tutti", "impegnarsi fino alle fine"), di abbandono (un tipo di sacrificio per gli altri o con l'altro o verso gli altri, o al loro posto, un sacrificio come avvicinarsi, ma anche come lavoro che chiede agli altri di unirsi, di ripetere la nostra azione e così di costruire un futuro di impegno comune) allo scopo di avvicinarci agli altri. Noi avanziamo (*engager*) o siamo avvicinati agli altri sia quando diventiamo legati a loro o li leghiamo a noi, sia quando "investiamo" in o "mettiamo qualcosa" prima degli altri.

Cosa significa ciò? Cosa significa mettere un pegno (garantire, dare in cauzione, ipotecare) prima di un altro o prima di tutti (l'intera comunità), e fino a che punto questa è una forma di modesta violenza che forza gli altri (o tutti) a scegliere se unirsi in questa specifica azione o meno? Quale tipo di azione non deve principalmente essere in stretta relazione con l'altro ("se sto facendo qualcosa, allora tu o lui dovete fare di conseguenza"), ma deve certamente legarmi all'altro e legare l'altro a me in modo tale da

obbligarci congiuntamente a porla in essere ("se io agisco, allora noi tutti agiamo"; "se tu agisci, allora tutti agiscono")? Se le mie attività pubbliche comportano la raccolta di fondi per prendersi cura dei bambini gravemente ammalati, organizzare rifugi temporanei per i rifugiati di guerra di uno stato vicino, o se visito spesso mattatoi per protestare contro quel modo di uccidere gli animali, non sarebbero tutte queste attività da chiamarsi "impegnate" (e "di attivismo")? Ognuna potrebbe rappresentare un impegno personale, e allo stesso tempo nessuna potrebbe essere eseguita individualmente, ma richiederebbe sempre un gruppo più o meno grande di persone, un impegno comune. Tuttavia, questa trasformazione dell'individuo in un agente di gruppo non ha necessariamente bisogno di essere considerata la più significante caratteristica di queste azioni. Kant ha inaugurato la spiegazione di questa trasformazione, lì dove parla del dovere verso se stessi in quanto tale, del debito od obbligazione a se stesso che precede sempre e sostiene/condiziona qualsiasi possibile obbligazione verso gli altri, ossia il dovere esterno.

Più complicato, ma forse più decisivo, è l'insieme di azioni che potrebbero essere localizzate in quel luogo dove la lingua inglese fa allo stesso tempo converge e divergere due parole o strategie complementari: *engagement* e *commitment*. L'azione d'impegno personale (forse crucialmente in contraddizione con il francese *engagement*) rimane personale, come nell'impegno per la mia carriera o la cura da una malattia. Solo una manciata di persone del mio ristretto circolo riconoscerà questo impegno, e nel riconoscerlo potrebbe sembrare loro essere "una cosa di pubblica importanza" e dunque un'obbligazione a unirsi. L'impegno nel senso di *commitment*, o l'impegno in comune, essendo sempre declinato al plurale, chiama in gioco un tipo differente di obbligazione. Quando organizzo un incontro a pranzo del mio gruppo in un ristorante, e prometto di partecipare all'inizio dell'incontro, allora sono veramente impegnato e coloro che rispondono alla chiamata per l'incontro confermeranno la mia azione, diventando così a loro volta impegnati. Ma l'impegno del nostro gruppo ("agire in concordanza con gli impegni") si verifica solo quando le azioni del gruppo producono una ragione sufficiente o una qualche forma di obbligazione per coloro che non appartengono inizialmente al nostro gruppo, o per coloro che non sono ancora nel programma dell'incontro, a unirsi necessariamente. Se il nostro gruppo agisce realmente in sintonia, insieme, se è impegnato in comune, allora sembrerebbe che io sia obbligato a unirmi, a diventare impegnato ("se tutti agiscono, allora io agisco"). Questa, che possiamo chiamare una nuova obbligazione, è differente da una obbligazione non-perfetta perché,

ad esempio, la persona che fa la carità non produce in alcun modo un'identica obbligazione in me. Per contrasto, l'impegnarsi di un gruppo non potrebbe mai lasciarmi indifferente.

Di conseguenza, vorrei aggiungere un nuovo tipo di "atto d'impegno", la cui caratteristica sarebbero d'essere compiuto da un gruppo (o è compiuto in un gruppo, come parte di un gruppo, ma, questo il punto cruciale, non esiste senza un gruppo, anche se la sua origine è possibile che risieda nello pseudo-istituto della Roma Repubblicana dello *ius provocationis*). Questi atti obbligano e connettono soprattutto tutti coloro che non sono ancora parte di un gruppo o di un'istituzione, che non lo sono ancora diventati, nonostante siano sempre presenti.

Il grido di "aiuto" obbliga coloro che non sono presenti, si rivolge a chiunque lo ascolti, anche se non è mai stato ascoltato prima. Né la sua forza imperativa e la sua capacità *ad hominem* sono ancora più deboli di come sarebbero se fosse una richiesta individuale, un grido o una supplica: "aiutatemi". L'impegno comune assume in primo luogo un atto pubblico che implica l'appartenenza a un gruppo, come quello di coloro che potrebbero aiutare ("Io agisco, di conseguenza sono un agente, una parte di un tutto, di tutti quanti insieme"). Solo se faccio qualcosa pubblicamente, rivolgendomi potenzialmente a tutti quanti e affermando l'esistenza di tutti quanti, solo allora io provo l'esistenza di un gruppo e la mia appartenenza ad esso.

Tutte queste condizioni implicano necessariamente la possibilità che l'impegno comune fallisca o che si verifichino degli insoddisfacenti atti d'impegno – ad esempio un gruppo incapace di incorporare nel suo impegno quelli che non gli appartengono. Ciò che confermerebbe l'esistenza di un mondo comune. Probabilmente solo allora sarebbe possibile parlare di atti sociali negativi, che, seppur transitori, rimangono in ogni caso una parte normativa degli atti sociali.

English title: What is an engagement act? Husserl and Reinach on 'Subject of the Superior Level' (We) and (Non) Social Acts.

Abstract

My intention is to describe one kind of "social act" that I have called "engaged act" (and which should be different from "joint commitment" although the English 'commitment' is often translated into German or French

as engagement*). I wish to uncover and demarcate these engaged acts in Husserl's endeavor to define and* de facto *establish social acts as such. My parallel tasks would be: to show the importance of social acts in the construction of some kind of new entities, which it is always problematic to name; to distinguish social acts as clearly as possible from empathy; to name some social acts "engaged acts" thus alleviating and clarifying Husserl's efforts in the course of determining social acts; and to re-evaluate Adolf Reinach's contribution to defining social acts in comparison with Husserl.*

Keywords: engagement; act; social act; commitment; we.

Petar Bojanić
Institute for Philosophy and Social Theory - University of Belgrade
bojanicp@gmail.com

Philosophy, Knowledge, and the Sciences

T

L'appartenance. Vers une théorie de la chair

Renaud Barbaras

Mon interrogation porte sur ce que nous nommons corps faute de mieux, autrement dit le corps propre, le corps vécu ou encore la chair. Ce terme traduit le *Leib* allemand, qui désigne le corps corrélatif du *Leben*, du vivre, à savoir le corps qui est indistinctement vivant et vécu. On pourrait dire d'ores et déjà que penser le corps dans sa singularité c'est découvrir un mode d'être qui soit neutre par rapport au partage du vivre intransitif et du vivre transitif, du *leben* et de l'*erleben*, de la vie organique et de l'expérience. La question que nous nous posons, à la suite de nombre de phénoménologues, est celle du sens d'être de cela que nous nommons corps, du mode d'existence qui le caractérise. Or, il y a une difficulté fondamentale à le penser, déjà déposée dans le terme même (mon corps : c'est un corps comme les autres et ce n'est pourtant pas un corps comme les autres puisqu'il est mien), qui se résume à ceci que nous sommes toujours en-deçà ou au-delà du point où il se situe, que nous le manquons à la fois par défaut en en faisant un simple fragment de matière et par excès en prétendant fonder sa différence sur la présence en lui d'un principe étranger au corps, âme, conscience ou esprit. Cette difficulté fondamentale renvoie à ce que Hans Jonas a nommé l'ontologie de la mort, qui domine l'ontologie depuis l'âge classique. Pour cette ontologie, la mort, c'est-à-dire la réalité inerte qui est celle du non-vivant, est la norme de tout ce qui est, ou encore le sens d'être fondamental vis-à-vis duquel la vie apparaît comme une exception. Cette ontologie vient converger avec le courant d'origine gnostique, qui tend à réduire toute vie à celle de l'âme et est donc en accord avec l'objectivisme naturaliste sur le fait que la vie a déserté le monde. Pour une telle ontologie, le corps vécu n'a aucune place dans le tableau de la réalité, il est en droit réductible à une machine couronnée d'une âme.

 DOI: 10.4454/teoria.v39i1.65

Cette perspective met fin et succède à une ontologie universelle de la vie pour laquelle le corps propre, c'est-à-dire l'unité psycho-physique (il faudrait dire l'identité) constitue au contraire la norme de tout étant. On voit d'ores et déjà, comme l'indique le terme, qui enregistre une continuité ontologique avec le reste de l'étant, que la question du corps a nécessairement une portée ontologique, qu'elle ne peut être résolue localement, bref qu'elle met en jeu le sens d'être de ce qui est.

Or, il n'est pas sûr que la question soit bien posée. Tout se passe comme si en parlant de corps on en avait déjà trop dit, comme si on avait déjà proposé une solution avant même d'avoir soulevé le véritable problème. Bien entendu, l'expérience dont il est question est évidente et pour ainsi dire originaire, même si elle est obscure, mais la question est de savoir si elle peut être caractérisée par le concept de corps et si ce concept nous permet de nous l'approprier. En vérité, le corps apparaît comme une réponse à une question qui n'a jamais été posée, comme la détermination d'un problème qui en occulte la problématicité. Il est donc nécessaire de mettre au jour ce qui est en question sous le concept de corps, le problème dont il est le nom. Plus précisément, le corps est un étant et il s'agit de savoir à quel mode d'exister ou à quelle expérience fondamentale cet étant renvoie. Cette expérience fondamentale est celle de *l'appartenance* : avoir un corps signifie en vérité appartenir au monde. Notons que c'est ce que signifie le recours à l'incarnation pour penser le corps puisque s'incarner c'est venir au monde, sauf que au monde nous y sommes toujours déjà. Si on renonce donc à projeter des catégories ontologiques non interrogées, il faut reconnaître que le corps n'est rien d'autre que cela en et par quoi j'appartiens au monde, de telle sorte que l'expérience du corps nous reconduit à celle de l'appartenance. Ainsi, ce n'est pas parce que j'ai un corps que j'appartiens au monde. Sous l'apparence de l'évidence, cette proposition est chargée de présupposés : elle engage un certain sens du corps comme fragment d'étendue, de l'appartenance comme inclusion objective (occupation d'une place) et du monde comme extension objective. Il faut donc affirmer au contraire que c'est dans la mesure où nous appartenons au monde que nous avons un corps et qu'avoir un corps ne signifie rien d'autre et rien de plus qu'appartenir. De sorte que la question du corps est tout entière reportée sur celle de l'appartenance, dont Louis Lavelle souligne à juste titre qu'elle est le fait primitif : « Mais le fait primitif, c'est que je ne peux ni poser l'être indépendamment du moi qui le saisit, ni poser le moi indépendamment de l'être dans lequel il s'inscrit. Le seul terme en présence duquel je me retrouve toujours, le seul fait qui est

pour moi premier et indubitable, c'est ma propre insertion dans le monde »[1]. On aperçoit déjà que s'il s'agit bien du fait primitif, le sens du moi comme celui du monde devront être abordés à partir de l'insertion elle-même, que l'appartenance commandera la nature des termes en présence.

Même s'il nous appartiendra d'en approfondir le sens, il faut déjà souligner que l'appartenance dont il est question ne se réduit pas à la dépendance logique de l'espèce au genre ou au fait d'être une unité dans un groupe additif. Il faut plutôt la comprendre à partir de l'appartenance à un courant ou un mouvement, pour autant que celle-ci enveloppe une forme de réciprocité : le mouvement est constitué par ceux-là mêmes qui lui appartiennent, de telle sorte qu'il leur appartient aussi. L'appartenance a donc le sens d'une participation ontologique : dire que j'appartiens au monde, c'est dire que j'en suis ou que je le suis, sans pour autant que je m'y dissolve, de sorte que c'est en l'étant que je suis moi-même, en lui appartenant que je m'appartiens. Ici, il n'y a pas d'alternative entre le même et l'autre, entre être soi-même et être autre chose que soi-même puisque c'est en étant du monde que je suis ce que je suis. Penser l'appartenance c'est précisément penser cette situation, c'est comprendre comment je peux, sous le même rapport, à la fois être moi-même et être autre que moi, moi-même en étant autre. Telle est en tout cas l'expérience même du corps, un être auprès de soi dans la dépossession, un entrer en soi qui est un sortir de soi. Comme le dit Alphonse de Waelhens, « le corps est ce qui nous fait être comme étant hors de nous-mêmes » Ceci appelle deux remarques. D'une part, il suit de cette première analyse que l'appartenance ne doit pas être comprise comme ce qui arrive à un je déjà constitué (on se demande d'ailleurs ce que pourrait bien signifier pour un je appartenir au monde) : c'est au contraire le je qui procède de l'appartenance comme fait originaire, ce qui revient à dire que le je est une dimension du corps loin que le corps soit ce que posséderait un je. Dire que j'ai un corps, c'est dire que j'appartiens mais dire que j'appartiens, c'est mettre au jour une inscription qui est la condition même de l'ipséité. D'autre part, si le sens de l'appartenance est celui d'une participation, en laquelle l'être-soi et l'être-autre ne font pas alternative, on soupçonne déjà que le degré de subjectivité du sujet pourra aller de pair avec la profondeur de son inscription dans le monde, que la possession (de soi par soi) se mesurera à la dépossession.

Il suit de là, au titre de position ontologique fondamentale corrélative d'une phénoménologie de l'appartenance, que tout ce qui est ou est sus-

[1] L. Lavelle, *De l'Acte*, Aubier, Paris 1937, p. 10.

ceptible d'être appartient au monde, y compris cela même que la tradition pouvait considérer comme lui étant le plus étranger, à savoir précisément le sujet transcendantal. Bref, rien ne peut être étranger au monde car le monde est l'omni-englobant universel : être c'est toujours *en* être ou *y* être. Autrement dit, cette réflexion sur le corps débouche sur une ontologie de l'appartenance dans la mesure où ce qui vaut pour lui, dont on pouvait considérer qu'il possédait une dimension étrangère au monde, vaut a fortiori pour tout étant. Dire que le monde est l'omni-englobant, c'est reconnaître que tout étant lui appartient et que la différence entre les étants ne peut alors consister qu'en leur modalité d'appartenance. A l'ontologie classique, pour laquelle la différence d'être fondamentale passait entre une appartenance à un monde nécessairement rabattu alors sur l'étendue, et une non-appartenance, celle du sujet ou de la conscience, il faut substituer une ontologie pour laquelle la différence passe entre plusieurs manières d'appartenir au monde puisque tout ce qui est *en* est. Autrement dit, ce n'est pas l'être mais bien l'appartenance qui se dit en plusieurs sens. Toute la question sera alors de mettre au jour un principe de distinction entre les appartenances. Le corollaire de cette ontologie de l'appartenance est une ontologie universelle des corps. En effet, si avoir ou être un corps signifie appartenir, si donc corporéité et appartenance se réciproquent et si tout étant est du monde il faut en conclure qu'il n'y a que des corps. Mais cela signifie aussi qu'il y a plusieurs manières d'être corps. Nous introduisons donc un principe de différenciation qui est transversal par rapport à celui de la philosophie classique : la prétendue différence entre ce que l'on appelle conscience et corps, qui est différence entre un hors-monde et un intra-mondain, renvoie en vérité à la différence entre des corps au sein du monde (ce qui est finalement tautologique), c'est-à-dire entre des modes d'appartenance. Être une conscience, c'est appartenir de cette façon singulière que l'on ressaisit à travers ce que l'on nomme corps propre (être conscient, c'est donc avoir un corps), être une chose c'est aussi appartenir au monde, mais d'une autre manière et, en vérité, de plusieurs autres manières car il y a plusieurs types de choses.

Il suit de tout cela que la *spatialité* est constitutive de l'être de tout étant. En effet, affirmer que l'appartenance est le mode d'être originaire de l'étant, c'est reconnaître que, pour tout étant, être signifie nécessairement *y* être. Appartenir au monde c'est y être situé, y occuper un lieu. En d'autres termes, la question « où ? » engage toujours la réalité même de ce qui est questionné, elle a nécessairement une portée ontologique et est, à ce titre, la question la plus profonde. Bien entendu, le sens de ce « où »

est commandé par le contexte ontologique dans lequel il s'inscrit. Il ne renvoie pas à une place au sein de l'espace objectif mais à un mode d'appartenance : la question concerne donc un type d'occupation ou d'habitation, un « comment ». C'est pourquoi notre perspective est aux antipodes de l'approche classique, qui est encore celle de la phénoménologie. Selon celle-ci, la différence du sujet avec le monde est différence avec des étants qui sont ce qu'ils sont (les substances) et elle implique donc que le sujet ne soit pas ce qu'il est : pour un sujet, différer du monde signifie différer de lui-même, c'est-à-dire exister temporellement. Dès lors, la situation dans le monde devient nécessairement extrinsèque, étrangère à l'essence du sujet, de telle sorte que la question de la spatialité devient par là-même secondaire. A l'inverse, en faisant de l'inscription dans le monde l'essence de tout étant et donc du sujet, on met la spatialité au cœur de celui-ci, étant entendu que cette spatialité est synonyme d'appartenance et se trouvera donc modalisée à partir de celle-ci. En d'autres termes, la spatialité telle que nous la comprenons d'emblée, spatialité objective au sein de laquelle l'étant occupe une place circonscrite, n'est plus qu'une modalité parmi d'autres de la spatialité, à savoir celle qui correspond à la simple chose, ou plutôt dont elle définit le mode d'être.

La question est alors celle du mode de spatialité des autres étants, pour autant qu'il ne se réduit pas à la place ; plus précisément, elle est de savoir qu'est-ce qui est pour les autres étants, et notamment notre corps, ce qu'est la place pour la simple chose. On comprend d'ores et déjà que l'espace dont il s'agit ici est un espace spatialisant, pour autant que chaque étant occupe l'espace à sa façon et déploie donc le type d'espace qui lui correspond : appartenir signifie non pas occuper une place mais spatialiser, c'est-à-dire déployer un lieu au sein de ce lieu de tous les lieux, de ce *où* originaire qu'est le monde. Mais on comprend aussi que, au sein de ce où, les différences entre modes de spatialisation ne pourront être que des différences d'ampleur ou de profondeur. Dire que les étants appartiennent au monde, c'est dire qu'ils ne sont pas seulement là où ils sont au sein de l'espace objectif, bref qu'ils excèdent leur place, qu'ils occupent le monde comme tel à leur façon, qu'ils empiètent sur lui ou le rassemblent. C'est en ce sens que Bergson distingue un corps intérieur et central et un corps immense, qui « va jusqu'aux étoiles »[2]. Il ne faut donc pas dire que je suis ici par mon corps et aux étoiles par ma vision, ce qui soulève plus de problèmes que cela n'en résout, mais que cette vision des étoiles suppose que

[2] Cfr. H. Bergson, *Les Deux Sources de la Morale et de la Religion*, Alcan, Paris 1932, p. 277.

j'occupe l'espace jusqu'à elles (tel est le sens véritable de la perception) et que, à ce titre, je possède un corps immense qui s'étend jusqu'à elles. Ce monde jusqu'aux étoiles, c'est là où *je* suis. De même, pour reprendre un exemple d'Augustin Berque, le lieu de cet outil qu'est un crayon n'est pas la place objective qu'il occupe comme fragment de matière car, précisément, le crayon comme tel n'est pas ce fragment de matière mais ce qui est impliqué par l'écriture qu'il rend possible et qui est comme le milieu du crayon. Ce mode d'être, qui n'est ni subjectif ni proprement objectif et que Augustin Berque nomme *trajectivité*, est le tissu relationnel au sein duquel le crayon existe et sans lequel il n'existerait pas, tissu qui est son lieu propre et qui excède sa place matérielle. Ainsi, en tant que modes d'appartenance, les étants ouvrent un type d'espace qui leur est propre, ils déploient ou constituent leur lieu ou leur sol d'appartenance, ils possèdent une forme d'ubiquité par rapport à l'espace objectif.

Reconnaître qu'il y va du lieu dans l'être de tout étant pour autant que tout être signifie un mode d'appartenance, c'est réaliser une forme de jonction entre l'ontologie et la géographie, c'est ouvrir la voie de ce que l'on pourrait nommer une ontologie géographique : la question de l'être renvoie à celle du *où*. C"est ce que souligne avec profondeur Augustin Berque, « 'Il y a ceci plutôt que cela et ici plutôt que là' : ce n'est pas seulement la géographie mais l'ontologie aussi que fonde un pareil énoncé – le premier énoncé, en vérité, que peut faire un être humain dès qu'il s'éveille à l'existence. Dire que la question de l'être est philosophique, tandis que celle du lieu, elle, serait géographique, c'est trancher la réalité par un abîme qui interdit à jamais de la saisir ». Et c'est pourquoi il ouvre l'ouvrage par l'affirmation suivante : « Il manque à l'ontologie une géographie et à la géographie une ontologie »[3]. Plus précisément, ce qu'Augustin Berque appelle de ses vœux est ce que nous pourrions nommer une géographie ontologique, géographie pour laquelle il y va de l'être dans le lieu, pour laquelle le *où* n'est pas extrinsèque mais commande la nature de ce qui y est. Or, cette géographie ontologique, qui donne tout son poids au lieu mais demeure une géographie, suppose bien une ontologie géographique pour laquelle il y va du lieu dans l'être, pour laquelle l'être de tout étant signifie un mode d'appartenance, ou encore d'occupation du monde et qui met au jour les conditions auxquelles il y va du lieu dans l'être. C'est cette ontologie que nous nous proposons de développer.

[3] A. Berque, *Ecoumène. Introduction à l'étude des milieux humains*, Belin, Paris 1987, p. 12, 10.

Afin de fonder la possibilité de cette ontologie, c'est-à-dire de modaliser l'appartenance, il est nécessaire d'en approfondir le sens. En effet, les différences au sein de l'appartenance ne peuvent être que des différences de profondeur ou de degré, même si elles se manifestent dans des modes d'existence résolument distincts. Or, une appartenance plus profonde ou plus radicale ne peut rien signifier d'autre que le fait, pour le sujet, d'être en relation avec « plus » de monde, de se rapporter à une plus grande extension du monde et d'être ainsi plus près du monde comme tel. *Appartenir* au monde, c'est appartenir au *monde*, de telle sorte que le niveau de l'appartenance se mesure au degré de présence du monde : tel étant appartient d'autant plus au monde qu'il *a* pour ainsi dire plus de monde. Ainsi, sa présence dans le monde est proportionnée à la présence du monde en lui ; sous le rapport d'être (dire que le sujet appartient au monde, c'est dire qu'il est *du* monde) se fait jour une relation qui est de l'ordre de l'avoir : être au monde, être dans le monde, ou encore être du monde c'est toujours avoir le monde. Mais il va de soi que l'appartenance du monde à l'étant ne peut être de même nature que celle de l'étant au monde. Comment en effet l'étant peut-il posséder plus de monde sinon en ouvrant à lui, en le faisant paraître? La présence de l'étant dans le monde est présence du monde dans l'étant, mais on pourrait dire que la vérité, qui est ontologique, de la première présence, repose dans la vérité, qui est cette fois phénoménologique, de la seconde. Si on veut bien se défaire de l'espace objectif et donc de la réduction de l'appartenance à l'inclusion spatiale, il faut reconnaître qu'un étant appartient au monde dans la mesure exacte où il se dilate à ses dimensions, c'est-à-dire encore ouvre à lui, le fait paraître. L'intramondanéité comme inscription ontologique est toujours cosmophanie.

Nous nous situons ici à un niveau qui n'est autre que celui de l'essence de l'intentionnalité – en un sens très large, qui convient en droit à tous les vivants et même à tous les étants dans le cadre de ce que Baptiste Morizot a nommé un animisme méthodologique. En effet, l'idée ici est que toute visée, tout rapport de connaissance suppose un rapport d'être, que sous l'acte par lequel je me rapporte à quelque chose il y a, comme sa condition de possibilité même, une proximité spatiale qui renvoie à une communauté ontologique : je ne peux viser quoi que ce soit que dans la mesure où j'y suis et où, par là même, je le suis d'une certaine façon. Ainsi, ce n'est pas parce que je vois les étoiles que je peux éventuellement les rejoindre par mon corps; c'est au contraire parce que mon corps va jusqu'aux étoiles, c'est-à-dire est ontologiquement joint à elles que je peux

les voir. Ou plutôt, plus rigoureusement, la possibilité de voir les étoiles tout comme celle de m'y joindre ou de les rejoindre sont les deux faces d'une même situation ontologique que je nomme précisément appartenance. Cette appartenance délivre manifestement le sens d'être du corps pour autant que, par lui, j'y suis (par exemple aux étoiles) à la fois et indissolublement perceptivement et spatialement. C'est en ce se sens précis que, comme Merleau-Ponty l'avait vu, le corps détient le secret de l'intentionnalité. Quoi qu'il en soit, visée intentionnelle et proximité ontologique sont les versants déjà abstraits de l'appartenance, qui est leur point même d'indistinction. Comme on le voit, cette appartenance peut également être définie par la réciprocité non-contradictoire de l'avoir : dire que j'appartiens au monde, c'est exactement dire qu'il m'appartient ; il ne m'a que si je l'ai. Ou encore, l'enveloppement ontologique de l'étant par le monde est nécessairement l'envers d'un enveloppement phénoménal du monde par l'étant : le monde ne me tire à lui que si je le tire à moi et inversement. Mais il faut comprendre ici que cette relation qui définit l'appartenance vaut pour tout étant et permet dans cette mesure de les discriminer. C'est pourquoi il faut sortir de la perspective, qui est encore celle de Merleau-Ponty par exemple, qui assignerait cette relation et donc l'intentionnalité au seul sujet, ou plutôt qui limiterait le sujet à la seule existence humaine. Nous avons affirmé que tout rapport de connaissance renvoie à un rapport d'être, que toute possession intentionnelle est l'envers d'une dépossession ontologique. Il s'agit alors seulement de poser la converse et de conclure que tout ce qui appartient au monde, c'est-à-dire tout étant en est nécessairement un mode de phénoménalisation : tout étant possédé par le monde possède celui-ci. Si toute cosmophanie est inscription, toute inscription est cosmophanie.

Il s'agit alors de mettre au jour les conditions sous lesquelles de telles affirmations sont pensables. Or, s'il est vrai que l'appartenance ne se confond pas avec l'inscription à une place mais signifie une occupation du monde qui en est une phénoménalisation, force est de reconnaître qu'elle a une signification nécessairement *dynamique* et non plus statique. En effet, abordée statiquement, l'appartenance ne peut rien signifier d'autre qu'une certaine situation en un emplacement déterminé. Inversement, on comprend que l'appartenance, telle qu'elle a été caractérisée, ne peut renvoyer qu'à une forme de mobilité fondamentale : appartenir ce n'est pas faire partie du monde (en être une partie), c'est y entrer, c'est y dessiner un certain espace ou encore l'habiter et c'est en vertu de cet investissement dynamique, ou plutôt sur ce mode que le monde peut se figurer au

sein de l'étant qui en fait partie. C'est autour du mouvement que l'enveloppement mutuel de l'étant et du sujet peut s'effectuer : par exemple, on comprend que les sujets que nous sommes peuvent appartenir au monde par leur mouvement, de telle sorte que le monde leur appartient, dès lors que ce mouvement est phénoménalisant et donc plus qu'un simple déplacement. Autrement dit, l'avoir du monde qui sous-tend l'être dans le monde ne peut que posséder le sens d'une appropriation dynamique; le voir est toujours agir. Ainsi, l'appartenance ne peut être disjointe de l'espace objectif et donc modalisée qu'à la condition d'être comprise comme avancée ou investissement. Elle est toujours un pas dans le monde qui est un pas vers le monde, selon une forme d'itération fondamentale, comme si en se déployant dans le monde elle demeurait néanmoins toujours sur son seuil : l'appartenance signifie l'identité absolue d'un aller-vers et d'un être-dans. Tel est le sens véritable de l'appartenance : une sorte de déhiscence par laquelle l'étant, quel qu'il soit, maintient un écart vis-à-vis de cela à quoi il appartient profondément, demeure différent dans son identité ontologique même et le fait par là-même paraître. L'identité entre l'inscription et la phénoménalisation se réalise donc comme différence ou recul de l'étant au sein même du monde, ou plutôt devant le monde, différence ou recul qui constituent son étantité même.

Mais il faut évidemment s'entendre sur le statut de ce mouvement qui est inhérent à l'appartenance. La tentation est de le réduire à un déplacement ou, en tout cas, de lui reconnaître cette dimension, même si, lorsqu'il s'agit d'un mouvement vivant et donc phénoménalisant, il s'avère irréductible à un mouvement local. Si elle semble phénoménologiquement fondée, une telle perspective a le défaut de maintenir la coupure entre le vivant et le non-vivant, coupure qui peut apparaître comme un dernier avatar du dualisme et interdit de faire place aux autres étants dans cette ontologie de l'appartenance. En vérité, qu'on le veuille ou non, maintenir une dimension de déplacement au sein du mouvement phénoménalisant c'est demeurer tributaire de l'espace objectif et, par conséquent, d'un sens limité de l'appartenance. Il faut donc aborder le mouvement à partir de l'appartenance au lieu de comprendre celle-ci, la dimension dynamique de celle-ci, à partir du mouvement comme déplacement. Le sens originaire du mouvement doit être recherché dans ce qui est effectué par l'appartenance, dans sa dimension proprement dynamique, qui consiste précisément dans l'ouverture d'un monde, dans la cosmophanie, plus précisément dans l'identité d'une habitation et d'une cosmophanie. Seul ce sens du mouvement, comme ouverture ouvrante, convient aussi bien aux déplace-

ments du vivant qu'au rassemblement d'un monde par une chose, une œuvre ou un paysage. Bref, ce n'est plus l'ouverture qui suppose notre mouvement mais au contraire notre mouvement qui est une modalité parmi d'autres de l'ouverture. Tous les étants relèvent donc du mouvement en tant que, appartenant au monde, ils initient à lui, même si seuls les vivants se déplacent.

Ces conclusions appellent au moins deux remarques. Tout d'abord, on pourrait nous accuser de tomber dans la circularité ou la tautologie puisque nous référons l'ouverture au mouvement pour définir ensuite le sens originaire du mouvement par l'ouverture. Mais cette circularité n'est problématique que d'un point de vue logique qui décompose le phénomène car, en vérité, elle se fonde dans l'unité même du phénomène. Parler de mouvement, ce n'est pas fonder l'ouverture sur autre chose qu'elle-même mais seulement en mettre en évidence une dimension. Dire donc que l'appartenance possède un sens dynamique c'est reconnaître seulement que l'ouverture qui est cœur de cette appartenance est une œuvre ouvrante et que c'est dans cette dimension que réside le sens le plus profond du mouvement – ce qui revient aussi à souligner que ce qui confère le statut de mouvement à nos déplacements est d'abord leur puissance d'ouverture, en quoi ils sont des gestes. La seconde remarque permettra peut-être de clarifier la situation. Si vraiment ce qui advient dans l'appartenance est une cosmophanie, c'est-à-dire l'avènement d'un monde, quelle qu'en soit l'ampleur ou la profondeur, alors le concept d'*événement* serait peut-être plus approprié. Avec l'appartenance, comme son sens même, quelque chose arrive (dans les deux sens du terme : se rapproche et se produit). Dès lors, ce que nous nommons mouvement, et qui renvoie au mode de phénoménalisation propre à certains vivants, ne serait qu'une modalité de l'événement. L'événement n'est pas ce qui vient infléchir un mouvement mais cela qui peut se réaliser ou non comme mouvement.

Ces résultats nous permettent d'en venir au statut du monde et d'éclairer en retour le sens de l'appartenance. Le monde est le sol originaire de tous les étants, le *où* ultime qui les unit tous, c'est-à-dire encore cela en quoi ils se trouvent et qui en eux se phénoménalise. Mais, comme on l'a vu, *y* être signifie toujours aussi *en* être, participer : appartenir c'est donc se distinguer de cela dont on est originairement fait ; l'appartenance signifie la différence au sein d'une unité ontologique première. Ce point est capital car il signifie que, en vertu de la communauté ontologique originaire sur laquelle s'enlève la différence, le mode d'être du sujet peut témoigner pour celui du monde, fonctionner comme échantillon ontologique du mode

d'être du monde. Autant dire que si, comme nous venons de le suggérer, le sujet ne peut appartenir qu'activement au monde et donc exister que sur le mode dynamique, on ne pourra se contenter d'une définition sommaire et statique du monde comme omni-englobant. Si le sujet peut y faire quelque chose, c'est dans la mesure où il *s'y* fait quelque chose, bref où le monde implique une processualité fondamentale. C'est sans doute à cette seule condition que l'on sera en mesure de dépasser définitivement le sens simplement inclusif de l'appartenance, qui implique inévitablement, qu'on le veuille ou non, de projeter le monde au plan de l'étendue. Le concept de *participation*, auquel nous avons cru pouvoir assimiler l'appartenance, prend ici toute sa portée. Participer ce n'est pas occuper une place dans un tout mais prendre part à quelque chose qui est en train de se faire. Comme l'écrit Lavelle : «La participation ne fait pas de nous, comme on pourrait le croire, une simple partie du Tout. Elle n'est pas une participation à un être déjà réalisé dont elle nous permettrait pour ainsi dire de nous approprier une part. On ne participe pas à une chose. On ne participe qu'à un acte qui est en train de s'accomplir, mais qui s'accomplit aussi en nous et par nous grâce à une opération originale et qui nous oblige, en assumant notre propre existence, à assumer aussi l'existence du Tout »[4]. Ainsi, comprise dynamiquement, l'appartenance au monde ne peut signifier qu'une participation au monde comme procès, à ce qui se fait dans le monde ou plutôt à ce que le monde fait. Mais le monde ne peut pas faire autre chose que ce qui se fait dans chaque étant, à savoir paraître comme monde, de telle sorte que le monde n'est rien d'autre que le mouvement de tous les mouvements, ce qui se fait en chacun d'eux, ou plutôt l'événement de tous les événements, l'apparaître qui est au cœur de toutes les apparitions. Il n'en est évidemment pas la simple somme puisque chaque étant fait paraître le monde à sa façon et renvoie donc à lui mais il n'est pas non plus autre chose qu'eux : il est ce qui confère l'unité à tous ces mouvements ou ces procès, cela qui en eux se phénoménalise mais résiste aussi à la phénoménalisation dans la mesure ou aucun ne fait paraître le monde même.

En ce sens, le monde peut être compris comme son propre procès de phénoménalisation et donc le cosmos comme cosmophanie originaire, étant entendu que l'infinité de ce procès implique un retrait irréductible du fond vis-à-vis de ses modes de manifestation, une inépuisabilité principielle. C'est pourquoi on peut dire tout autant que le monde est ce qui

[4] *Op. cit.*, p. 175.

résiste à tous les événements individuels de phénoménalisation, cela qui se manifeste à et en eux comme irréductible à eux. On comprend alors mieux, au vu de cette détermination dynamique du monde comme sa propre venue à la lumière, la réciprocité qui est au cœur de l'appartenance. Non seulement l'appartenance au monde n'exclut pas la phénoménalisation mais elle ne peut s'accomplir que comme telle puisque cela à quoi chaque étant appartient est précisément un procès cosmophanique, une incessante venue à la lumière. Appartenir au monde, c'est-à-dire participer à son œuvre, ne peut signifier rien d'autre que le faire paraître puisque tel est le mode d'être premier et ultime du monde. Si le monde est un procès phénoménalisant, descendre en lui signifie monter vers la lumière, s'y enraciner signifie le montrer. Ressaisie du côté du monde, la situation est celle d'un procès de phénoménalisation qui s'accomplit sous la forme d'étants individués, de telle sorte que chacun l'exprime à sa façon mais qu'aucun ne l'épuise. Nous sommes ici au plus près d'une sorte de relève phénoménologique du leibnizianisme.

Il faut en tirer une dernière conséquence, qui va à l'encontre de ce que la tradition a toujours présupposé. Pour celle-ci, il y a une relation d'exclusion mutuelle entre l'appartenance et la phénoménalisation, entre le corps et la subjectivité, si l'on veut utiliser le vocabulaire de cette tradition. C'est cette incompatibilité que concentre le mode d'existence du corps propre : d'un côté, il appartient au monde, en quoi il est corps à l'instar des autres corps ; de l'autre, il le fait paraître, en quoi il est le mien. Comment cela qui est du côté du monde, inscrit en lui, peut-il en avoir en même temps une expérience? Cela demeure incompréhensible Or, le concept d'appartenance nous a permis de renverser la situation pour autant que, comme participation, elle n'exclut pas mais exige une phénoménalisation. Telle est la leçon d'une phénoménologie de l'appartenance : il n'y a de parution du monde que du sein du monde, toute expérience du monde implique une inscription en lui; la communauté ontologique et la différence phénoménologique sont l'envers l'une de l'autre. Mais il faut en tirer la conséquence quant aux différents modes d'appartenance et donc aux différents étants : si l'épreuve du monde est l'envers d'une appartenance, celle-là est d'autant plus aiguë que celle-ci est profonde. Autrement dit, la puissance de phénoménalisation du monde, ou encore d'expérience, qui nous définit comme sujets, est proportionnée à la profondeur de notre inscription en lui. C'est donc parce que nous sommes du monde en un sens plus profond que les autres étants que nous sommes capables de le faire apparaître comme ne le fait aucun de ces étants. A l'inverse et

en quelque sorte à l'autre bout de la chaîne des appartenances, la simple chose inerte n'appartient au monde que superficiellement – ce qui signifie qu'elle n'est que là où elle est, située dans l'espace, identique à elle-même et donc simple substance – et, dans cette mesure, elle ne le fait paraître que minimalement, à savoir sous la simple forme de cela qu'elle est. Les pierres, comme le dit R. Caillois, « n'attestent qu'elles »[5]. Mais, pour ce qui est du sujet, la distance au monde qui en qualifie le sens d'être propre est l'envers d'une proximité radicale : c'est parce que le monde nous a plus profondément que les autres étants que nous l'avons plus radicalement qu'eux. C'est cette double possession qui est concentrée dans le mystère du corps. Ainsi, conformément à l'essence de l'appartenance, nous sommes loin du monde parce que nous en sommes très proches, nous le possédons parce qu'il nous possède : le face-à-face entre ce que l'on nomme sujet et objet est l'envers d'une connivence ontologique radicale. Bref, la différence phénoménologique inhérente à la corrélation est l'envers d'une communauté ontologique. Tel est le sens d'être véritable de la chair, qui n'est en vérité ni corps ni conscience mais une possession (perceptive) du monde qui est la contrepartie d'une dépossession (charnelle) par ce monde, une expérience du monde depuis sa profondeur, un être auprès de soi par un être au cœur du monde, une auto-affection par une altération ontologique. Tel est le pas que Merleau-Ponty ne franchit pas : ma chair appartient à la chair du monde mais c'est exactement en et par cette appartenance qu'elle est ma chair, loin que sa mienneté la distingue de celle du monde, comme celui-là le pensait encore.

English title: Belonging. Towards a theory of flesh.

Abstract

*This paper aims at showing that our lived body (*chair, Leib*), which is one of the most important topics of phenomenology, is not so much a question as an answer, answer to a question that remains implicit: that of* belonging. *It is not for having a body that we belong to the world; on the contrary, we have a body in so far as we belong to the world. Moreover, if the world is that which contains everything, in such a way that an existence out*

[5] R. Caillois, *Pierres*, Gallimard, Paris, *Dédicace*.

of the world is meaningless, we must conclude that any being belongs to the world and that the difference betweeen beings refers to their way of belonging. But, in so far as any belonging involves a ground, a philosophy of belonging leads out into a phenomenology of space, that distinguishes as many spaces as ways of belonging. Accordingly, the problem raised by the lived body is the following: what is the way of belonging of that being thanks to which the world itself appears?

Keywords: body; belonging; ontological geography; movement; participation; space; world.

Renaud Barbaras
Université Paris 1 Panthéon-Sorbonne
renaudbarbaras@orange.fr

T

Le "problème" des mathématiques

Jean-Michel Salanskis

Il n'y a pas véritablement accord sur ce qu'est la philosophie des sciences : il existe plusieurs façons de la définir et de concevoir sa mission. Cette pluralité d'acceptions, bien sûr, est liée à la pluralité des manières de comprendre la philosophie elle-même en général, mais elle n'en dépend pas de manière parfaite : des façons différentes d'appréhender la philosophie des sciences sont susceptibles de vivre juxtaposées à l'intérieur d'une sensibilité partagée au sujet de l'exercice philosophique.

Le point que je voudrais mettre en relief, et que je voudrais tenter de commencer à discuter, est que les mathématiques posent au moins tendanciellement un problème à presque toute approche de philosophie des sciences.

Ce point est paradoxal au moins à deux égards :

1) D'une part les mathématiques semblent à première vue simplement une des sciences dont une philosophie des sciences doit s'occuper. Comment et à quel titre poseraient-elles problème à la démarche elle-même ? Le pouvoir de nuisance maximal d'un objet vis-à-vis d'une enquête portant sur lui n'est-il pas son impénétrabilité ? Or, ce que nous allons tenter de montrer, c'est que le cas des mathématiques est susceptible de déstabiliser la démarche de la philosophie des sciences d'une manière en substance "plus radicale".
2) D'autre part, les mathématiques sont regardées, depuis fort longtemps, comme le lieu de la réassurance la plus grande pour la rationalité, pour la science. Il est pour le moins inattendu, dans ces conditions, qu'elles fassent problème pour une philosophie des sciences cherchant, du moins on peut le supposer, à s'approprier, refléter ou confirmer la rationalité des sciences.

 DOI: 10.4454/teoria.v39i1.66

Mais le "problème des mathématiques" n'est pas seulement une surprise ou un dérangement paradoxal pour la philosophie des sciences, il est aussi révélateur de l'essence des mathématiques, ainsi que, d'ailleurs, de la place des mathématiques dans la culture. La présente réflexion prend donc ce problème comme une chance, elle y voit un bon angle pour déchiffrer quelque chose du mystère des mathématiques, une incitation salutaire à prendre les mathématiques au sérieux et accueillir leur performance.

Essayons, donc, d'entrer dans ce problème prometteur.

1. *Philosophie des mathématiques ?*

Une première face du problème est liée au projet d'une philosophie des mathématiques.

On considère le plus souvent, à ce qu'il me semble, que les notions de spécialisation de la philosophie par un objet appartiennent à l'époque contemporaine ou quasi-contemporaine. Qu'elles sont liées à une cartographie de la philosophie par elle-même largement fonction de la construction académique de celle-ci (Pierre Macherey observait, je m'en souviens, que Kant était le premier philosophe de l'histoire ayant travaillé dans le statut de professeur de philosophie à l'Université). En particulier la notion de philosophie des sciences est sans nul doute récente.

Pourtant j'ai entendu soutenir que la notion de philosophie des mathématiques est plus ancienne. Tout, ici, est affaire de définition, et je ne sais pas s'il faut valider un tel jugement. Certes, il est sans doute vrai que le champ philosophique accueille en lui avant la période récente des débats que nous reconnaissons irrésistiblement, dans l'après coup, comme des débats de philosophie des mathématiques : tel, typiquement le débat sur le calcul infinitésimal. Mais en même temps, chez Kant encore pour revenir à cet exemple, les considérations de philosophie des mathématiques que nous trouvons dans son œuvre ne sont pas concentrées et localisées dans des ouvrages dédiés de philosophie des mathématiques, ce qui peut sembler une condition pour que la branche puisse être dite exister.

Cela dit, Kant soulève dans au moins un texte le problème du rapport de la philosophie aux mathématiques : dans le court mémoire « Comment introduire en philosophie le concept de grandeur négative ? ». Ce texte est important pour nous en raison de son étrange posture : si Kant déclare, au début du mémoire, que le projet d'écrire la philosophie *more geometrico* est vain, et s'il souligne une sorte de disparité entre les deux disciplines,

interdisant en principe la moins sûre (la philosophie) de reprendre en son sein les résultats comme la méthode de la plus sûre (les mathématiques)[1], néanmoins il s'engage exactement dans le projet d'une telle reprise. Et au fil de sa réflexion, il en vient à affirmer une supériorité traditionnelle de la philosophie : le fait qu'elle sait mieux arrêter le contour de ce dont on parle, s'agît-il de contenus mathématiques[2].

Ce texte de Kant, d'ailleurs, semble se tenir en deçà de ce qu'on peut appeler philosophie des mathématiques : ce dont il traite et dans quoi il s'engage n'est pas une évaluation philosophique des mathématiques, il est plutôt question de réussir une hybridation en faisant venir quelque chose des mathématiques dans la philosophie (la notion d'opposition réelle).

Comme, par ailleurs, Kant a remarquablement pris le chemin de la philosophie des mathématiques (à nos yeux au moins) dans l'ensemble de son œuvre, notamment dans la *Critique de la raison pure*, nous voyons surgir à partir de lui un premier versant du problème. Ce serait celui-ci : le rapport des mathématiques avec la philosophie est-il entretenu de manière optimale par une philosophie des mathématiques, ou la démarche consistant à emprunter en philosophie quelque chose des lumières de la mathématique est-elle plus dans l'essentiel ?

Ce problème se pose de façon constante, et donne lieu à un dégradé subtil et passionnant d'attitudes. Un auteur comme René Thom, en France, a récemment donné des lettres de noblesse à l'idée d'une philosophie mêlant sa voix aux mathématiques, tirant parti d'elles[3]. Un auteur comme Jean-Toussaint Desanti, dans la même période, incarnait en revanche l'idée d'une gnoséologie hétérogène à la mathématique prenant intérêt pour elle, s'appliquant à elle. Même si Desanti tenait quant à lui à préciser que, dans une telle démarche, la philosophie devait pour une part parler depuis l'intérieur des mathématiques, en se refusant la facilité avec laquelle les philosophies traditionnelles avaient prétendu intérioriser à elles la mathématique[4].

[1] Cf. E. Kant, *Essai pour introduire en philosophie le concept de grandeurs négatives*, trad. franç. J. Ferrari, in *Kant Œuvres Philosophiques I*, Pléiade, Gallimard, Paris 1980, pp. 261-302 ; édition originale 1763. Ce que je résume dans ce paragraphe est le contenu du début de l'avant-propos (p. 261 ; II, 167).

[2] C'est le contenu, à peine explicite, du paragraphe de la page 264-265 [II, p. 170].

[3] Cf. R. Thom, *Modèles Mathématiques de la Morphogenèse*, Unions Générale d'Editions, Paris 1974.

[4] Cf. J.-T. Desanti, *Sur le rapport traditionnel des sciences et de la philosophie*, in *La Philosophie silencieuse*, Le Seuil, Paris 1975, pp. 7-109.

Une mention spéciale est due, à cet égard, à la démarche de la philosophie analytique. Par un aspect, ce qu'elle a fait, en révolutionnant le travail philosophique, c'est tout simplement le contraindre à suivre les rails et les voies argumentatives de la logique des prédicats du premier ordre, que des gens comme Frege avaient extraite en tant que forme de l'activité mathématique (typiquement l'arithmétique). Mais en fin de compte, la philosophie analytique a engendré une discussion philosophique "logico-linguistique" restant à distance du foisonnement et de la complexité du développement mathématique (même de ceux de la logique, à vrai dire). Et le nouveau mode philosophique a finalement produit une "philosophie des mathématiques" évaluant de l'extérieur l'édifice mathématique, comme on le fait en mode "continental". La sorte d'analyse que l'on conduit, dans ce contexte, est largement fondée sur la logique. C'est comme si la philosophie se glissait dans les vêtements techniques de la logique pour juger des mathématiques, en tirant argument du fait que la discipline logique semble avoir quelque droit à se dire en position fondationnelle par rapport à l'exercice des mathématiques. Néanmoins, sous la plume des auteur(e)s majeur(e)s du courant, on voit surgir des arguments proprement philosophiques, nous confirmant que la veine "philosophie des mathématiques" – au sens de l'application à l'évaluation ou l'interprétation des mathématiques d'une gnoséologie externe, proprement philosophique – n'est pas morte.

En même temps, si l'on s'en tient à la conversion de la philosophie à un régime argumentatif strict d'une part, à l'emploi de la consignation symbolique des formules d'autre part, on peut avoir le sentiment d'une philosophie engagée sur la voie de d'une nouvelle hybridation, de type méthodologique. La tension que nous essayons de décrire se maintient donc.

2. *La question de la physique*

De manière contemporaine, nous disons "Philosophie des sciences" en pensant de manière générale à une application de la philosophie aux sciences – à chacune des sciences, tour à – qui détermine en même temps une spécialité académique.

Pourtant, il est clair, si l'on examine les grands précédents fondateurs que sont, pour la philosophie des sciences, les démarches à certains égards programmatiques de Kant et de Carnap, que parmi les sciences, la physique est privilégiée.

Chez Kant, ce privilège est ouvertement avoué. D'une part, il décrit l'entreprise critique comme largement suscitée par l'admiration éprouvée devant la science newtonienne, allant, dans les *Prolégomènes…*, jusqu'à tenter d'exposer la critique à partir du fait épistémologique de cette science, au moyen d'une analyse régressive remontant d'un tel fait à ses conditions de possibilités[5]. D'autre part, dans les *Premiers principes d'une philosophie de la nature*, il explique dans la préface pourquoi seule la physique, à ce jour, est à proprement parler une science à ses yeux[6].

S'il en va ainsi, on le sait, c'est pour ainsi dire “à cause des mathématiques”, ce qui nous ramène à notre sujet. Pour Kant en effet, une science n'en est une que si elle est une connaissance a priori : si elle va au-delà du bilan des expériences, même augmenté de lois inductives. Une science doit énoncer des prédictions sur le mode d'une anticipation du monde qui en est, au fond, une reconstruction. Ce qui veut dire qu'elle “construit” un schéma du monde dans les intuitions pures de l'espace et du temps. Mais faire cela, c'est entrer dans la démarche de la construction de concepts qui est celle des mathématiques. Pas de science, donc, sans élaboration mathématique a priori de la région du réel dont on s'occupe. Il se trouve que, jusqu'à présent, seule la physique a fait cela, projetant sur le mode mathématique le concept empirique de “matière en mouvement”, ainsi que Kant s'attache à l'expliquer et l'exposer plus complètement dans les *Premiers principes*…. Kant juge que la psychologie ne pourra jamais le faire, et tente d'esquisser la manière dont la chimie pourrait y parvenir, prophétisant si l'on veut la chimie quantique[7].

La mathématique intervient donc, dans une telle approche, comme ce qui signe la scientificité de la science.

Le cercle de Vienne, venant après, fut impulsé par des auteurs nourris de culture kantienne. On trouve chez Carnap, le père fondateur de la nouvelle école, de fréquentes allusions à la philosophie kantienne. Le décalage, de la philosophie kantienne de la science, vers une philosophie analytique des sciences d'une espèce nouvelle, se fait d'ailleurs de manière lente et problématique, pour qui regarde ce développement avec un regard d'historien. Peut-être même la strate kantienne n'est-elle jamais complète-

[5] Cf. E. Kant, *Prolégomènes à toute métaphysique future qui pourra se présenter comme science*, trad. franç. J. Rivelaygues, Éditions de la Pléiade, sous la direction de F. Alquié, tome II, Gallimard, Paris 1985, pp. 17-172 [IV, pp. 255-383] ; édition originale 1783.

[6] Cf. E. Kant, *Premiers principes métaphysiques de la Science de la nature*, trad. franç. J. Gibelin, Vrin, Paris 1982, pp. 11-13 [IV, pp. 470-471]; édition originale 1786.

[7] Cf. *Premiers principes métaphysiques de la Science de la nature*, cit., p. 12 [IV, pp. 470-471].

ment abandonnée, si l'on en juge par des "come back" impressionnants et récurrents[8].

En substance, à l'époque carnapienne, la science est encore conçue comme une manière de "sauver les phénomènes", elle est donc originairement décrite comme incluse pour ainsi dire dans une "réduction aux phénomènes" que la jeune philosophie analytique va pourtant, de plus en plus, reprocher à Kant. La différence, d'emblée décisive, est néanmoins que l'information provenant de l'expérience est supposée toujours traduite dans des *énoncés protocolaires*, comptes rendus de ce qui se passe dans le laboratoire. Ces énoncés protocolaires, on le sait, seront ce par rapport à quoi on vérifie les lois universelles se déduisant de la théorie scientifique courante, supposée mise dans la forme exemplaire et typique d'une théorie logique du premier ordre. Mais cela suppose que l'on admette, au niveau de la formulation de ces énoncés protocolaires, une référentialité simple et non problématique. Ce qui nous permet de prendre $R(t_1,\ldots t_n)$ comme une entrée informative provenant de la nature, c'est que dans le contexte du laboratoire nous saisissons les objets dénotés par $t_1,\ldots,t_n$ et sommes en mesure de constater à leur propos la propriété R. L'empirisme logique en train de naître est donc amené à rompre avec le kantisme en se donnant la référentialité simple du langage comme base pour toute l'élaboration scientifique, au lieu de prendre l'objet, déjà pour la connaissance ordinaire, comme le fruit d'une synthèse, toujours à reprendre et à remettre en cause dans la science. Telle est la manière d'hériter de la révolution référentialiste opérée par Frege et Russell.

Il en résulte que la question des mathématiques suscite une sorte d'antinomie fondamentale de la philosophie des sciences, que je résumerai dans les termes suivants :

1) D'un côté, nous ne pouvons que prendre acte de ce que le mode de la physique mathématique est celui du triomphe contemporain de cette science, et cela nous forcerait à interpréter la physique comme Kant.
2) De l'autre côté, si nous sommes convaincus par Frege et Russell, nous ne pouvons comprendre la science que comme une élaboration théorique fondée sur les liens référentiels fondamentaux de notre langage avec le réel.

[8] Tel celui de John McDowell (cf. J. McDowell, *L'esprit et le monde*, trad. franç. C. Alsaleh, Vrin, Paris 2007).

Si l'on regarde la science côté 1), on la voit comme caractérisée par son usage des mathématiques. Un usage qui consiste à projeter a priori le réel sur la toile d'une structure mathématique, ce qui conduit d'entrée de jeu à une rupture avec le sens commun. Puis, au-delà, à une faculté de rupture de la science avec elle-même, rupture obtenue à chaque fois par l'adoption d'une nouvelle structure mathématique de référence (on échange le $\mathbf{R}^3$ euclidien contre une variété pseudo-riemanienne, ou l'un et l'autre contre un espace de Hilbert muni de quelques opérateurs auto-adjoints compacts). On interprète le rapport à l'empiricité aussi dans cette ligne, en soulignant que l'espace de repérage choisi dicte le recueil de données dans le monde (extraction d'objets mathématiques de l'expérience), et commande symétriquement la possibilité d'introduire des objets mathématiques dans le monde (préparation d'une expérience).

Si l'on comprend bien ces dimensions culturellement essentielles de la science, en revanche, on a de la difficulté avec l'autre critère de démarcation plus volontiers mis en avant dans la tradition empiriste : celui de la confrontation avec l'expérience, garantissant, notamment, le caractère réfutable des cadres mathématiques introduits. Ce qui fonctionne comme a priori mathématique, clairement, doit aussi être révisable (même si cette révision doit consister en une ré-anticipation, en une reprise de l'imagination mathématique du monde). Pour avoir une telle réfutation expérientielle, il semble bien qu'il faille, en fin de compte, que nous puissions formuler les résultats d'expérience comme des énoncés protocolaires: nous devons faire appel à la vérification du sens commun, fondée sur la référentialité frégéenne présupposée du langage.

Si l'on examine la science du côté 2) [post-frégéen], c'est-à-dire en substance avec un regard d'empiriste logique, alors on doit rendre compte du fait que la science reine peuple son discours d'entités n'ayant pas du tout de statut référentiel ordinaire. Le simple fait d'évoquer des vitesses instantanées place déjà le discours au-delà de la référentialité usuelle, puisque nous savons seulement définir une telle vitesse comme nombre dérivé d'une fonction, ce qui mobilise les objets idéaux de la mathématique. On a donc le paradoxe d'une science présumée réaliste – et célébrée comme la plus puissante à l'égard du réel – dont la démarche consiste néanmoins à raconter une histoire où tiennent une place centrale les objets problématiques, abstraits ou idéaux, de la mathématique. Les réponses classiques de la philosophie analytique des sciences, au premier chef de l'empirisme logique, consistent à expliquer cette couche mathématisante de la physique comme un surplus inessentiel par rapport à un

discours strictement empirique donnant lieu, via cette couche supplémentaire, à des prédictions elles-mêmes strictement empiriques (une possibilité typique consiste à supposer que la théorie mathématisante est une extension conservative de la théorie empirique).

Une dernière remarque, pour souligner le poids et la profondeur de ce "problème des mathématiques" dans la philosophie des sciences contemporaines. Est apparu, assez récemment, un argument de philosophie des mathématiques célèbre, que l'on baptise argument d'indispensabilité. On l'attribue à la fois à Quine et à Putnam.

Dans sa formulation la plus simple, il dit en substance ceci : nous sommes gênés vis-à-vis des mathématiques, parce que nous ne semblons pas pouvoir les considérer de manière confortable et plausible comme des sciences disant le vrai sur un objet externe. En revanche, nous voyons dans la physique par excellence le discours vrai par rapport au monde effectif. S'il est avéré, maintenant, que la mobilisation des mathématiques par la physique est incontournable (que les mathématiques sont *indispensables* à la physique), alors nous pouvons en conclure que les entités mathématiques sont elles aussi réelles. Quine soutient que dans les quantifications de nos discours résident nos engagements ontologiques : sont réputés exister par nous exactement tous les objets se laissant substituer aux variables sur lesquelles nous quantifions. Or le discours de la physique quantifie sur des objets mathématiques (pour toute valeurs réelles t et x du temps et de la position, $x=1/2gt^2+v_0t+x_0$ donne l'abscisse du point en chute libre). Donc ces objets "existent".

Le frappant est que cet argument renverse simplement l'ancien argument kantien : Kant disait plutôt que la mathématique enveloppait de la connaissance synthétique a priori, et que la science newtonienne, visiblement, entrait dans une formulation géométrique lui permettant à elle aussi l'énoncé d'un savoir synthétique a priori. Mais dire que la physique tient un discours synthétique a priori, c'est dire que nous ne pouvons pas la comprendre comme un discours empirique (elle serait strictement a posteriori).

Donc, la physique rend la mathématique réaliste pour Quine ou Putnam, alors que la mathématique rend la physique non strictement réaliste pour la tradition kantienne. Le même attelage est utilisé de deux manières contradictoires, et tout tourne autour du problème des mathématiques.

3. *Le computationnel*

Les mathématiques interviennent d'une nouvelle manière, au cœur de quelque chose qui possède à la fois une dimension épistémologique et une dimension "civilisationnelle" si l'on veut : je veux parler de ce que l'on peut appeler la révolution informationnelle.

Donc, ce dont il s'agit, c'est des machines calculantes, des ordinateurs, comme on les appelle le plus souvent : de la discipline nouvelle ayant nom *informatique*, et de la mutation polymorphe de la vie sociale qui résulte de la miniaturisation des machines ; de l'apparition du réseau mondial des réseaux, du progrès de l'informatisation des activités et des sciences.

Tout cela, qui s'impose à tout sujet de nos jours, qui ne laisse personne indemne, commence apparemment dans les mathématiques. La généalogie la plus commune et la plus évidente de la révolution informatique la fait remonter à l'article de Turing dans lequel il introduit la machine portant aujourd'hui son nom[9] : une "machine théorique", notion nouvelle et surprenante qui voit le jour en même temps. Ce qui s'ouvre à cette occasion est un compartiment de la logique que l'on appelle théorie de la calculabilité, ou une discipline nouvelle que l'on appelle informatique théorique. L'une ou l'autre sont liées aux mathématiques, en plusieurs sens :

1) La théorie de la calculabilité s'insère dans la logique contemporaine, que l'on désigne sous le nom de *logique mathématique*, au titre du fait qu'elle partage avec la mathématique sa méthodologie symbolique et déductive.
2) L'invention de la machine de Turing s'inscrit en même temps dans l'histoire de la théorie de la démonstration, qui elle-même "procède" au fond de la volonté de justifier la mathématique formelle: de ce que l'on a appelé le "programme de Hilbert". L'article de Turing se présente en effet comme une réponse à l'*Entscheidungsproblem* soulevé par Hilbert. Les mathématiques sont donc concernées du côté de leurs fondements.
3) L'informatique, nouvelle discipline, avec l'algorithmique comme sous-discipline, correspond en même temps pour une part importante d'elle-même à une "strate" de la mathématique, la strate des problèmes et des objets combinatoires, finis et discrets. L'informatique met en relief à l'intérieur des mathématiques une base objective (celle des objets finis et discrets en substance) et une forme fondamentale (celle du calcul) qui tiennent une place centrale et essentielle en elles.

[9] Cf. A.M. Turing, *On computable numbers, with an application to the Entscheidungsproblem*, in « Proc. London Math. Soc. », Ser. 2-42, 1936, pp. 230-265.

En tout cas, chacun reconnaît de nos jours que la philosophie devrait trouver ce qu'elle a à dire sur un tel sujet. On le lui demande même au nom de cette urgence sociale et politique à laquelle si souvent, on a voulu la ramener voire la soumettre. Et à laquelle d'ailleurs, depuis le cas du marxisme, elle s'est souvent abandonnée avec enthousiasme.

Or je pense que tout effort pour traiter de la révolution informationnelle à un niveau philosophique retrouve, pour ainsi dire sur son chemin, le problème des mathématiques. Je voudrais juste donner une idée résumée et synthétique de quelques aspects de la réflexion à laquelle nous engage la mutation informatisante à laquelle nous assistons. On trouvera une exposition plus complète et plus conséquente de ces matières dans mon *Le monde du computationnel*[10].

Il y a, d'abord, la question du sens en lequel on peut parler de *révolution*. Bien que nous disposions déjà d'usages hétérogènes du mot, puisque nous parlons facilement de révolution politique, de révolution économique ou industrielle et de révolution scientifique, il semble que jusqu'ici ces emplois trouvent une certaine unité dans la notion de "réinstitution" : nous parlons de révolution chaque fois que nous croyons observer, à un moment et dans un domaine, le passage de certaines formes d'organisation admises comme ayant autorité à de nouvelles formes. La révolution consiste justement dans l'abandon et le désaveu des anciennes formes au profit des nouvelles : un abandon et un désaveu dont on suppose de plus qu'il s'opère de manière systématique et volontaire. C'est pourquoi on peut parler de réinstitution : le mode délibéré, volontariste du geste instituant fait partie de la notion. Les cas des révolutions économiques ou industrielles est le moins net à cet égard, mais il reste un bon cas : la révolution électrique, par exemple, passe sans nul doute par des stratégies d'entreprises, ainsi que par une connivence délibérée de l'État et des entreprises.

De son côté, la révolution informationnelle consiste dans l'adoption d'un nouveau support (le support numérique) qui est à la fois support du recueil des données et lieu incontournable des traitements computationnels (des calculs). Ce nouveau support n'est pas institué comme le seul désormais légitime, destiné à remplacer les autres : dans beaucoup de cas il se présente à côté des anciens supports, permettant des actions en parallèle avec les anciennes actions. La déferlante du mode computationnel apparaît comme une tendance lourde que personne ne décide ou n'institue : plutôt, les institutions tentent de légiférer ex-post une fois que

[10] Cf. J.-M. Salanskis, *Le monde du computationnel*, Encre marine, Paris 2011.

le nouveau format des données et les nouvelles modalités des pratiques ont – déjà – changé le paysage.

L'apparition d'un tel nouveau support est sans doute quelque chose de plus radical que les anciennes révolutions, de plus absolument transversal à toutes les dimensions de la vie humaine. Et en même temps c'est tout à fait autre chose qu'une réinstitution : la meilleure comparaison possible serait avec l'invention de l'écriture, ou celle de l'imprimerie.

Les mathématiques sont concernées par l'apparition de ce nouveau support essentiellement à l'endroit suivant : le support numérique révèle la connivence et l'homologie profonde des formes linguistiques et des formes mathématiques. Que nous ayons pu entrer tous dans l'usage des traitements de texte, et que notre manière d'écrire s'en soit trouvée profondément transformée, cela “dérive” en un sens du fait que les expressions linguistiques se laissent coder, sans violence aucune, dans le champ numérique, et que nombre des opérations qui ont cours au niveau du texte se laissent ramener à la forme de la manipulation syntaxique, relevant en dernière analyse de la catégorie du calcul. L'étrangeté de la lettre et du nombre n'est plus pensable : l'informatique affiche publiquement une homologie de l'arithmétique et du linguistique que la pensée constructive avait mis au jour dans le débat sur les fondements, dès Brouwer sans doute.

Un second élément important est que nous sommes appelés à distinguer, dans la révolution informationnelle, ses deux niveaux d'opération. D'un côté, elle nous engage dans un mouvement de traduction de toutes les données partagées – de toutes les données de la culture – vers le format numérique permettant leur stockage, et leur affichage sous des formes variées sur les écrans servant d'interfaces aux outils numériques (ordinateurs, smartphones, tablettes). De l'autre côté, elle nous permet de concevoir au plan numérique des actions portant sur les données : des “calculs” au sens de Turing, Gödel et Church. Ces actions, parfois, suppléent à des actions que nous accomplissions autrement, directement sur les données dans leur état non numérisé. Dans d'autres cas, elles sont des actions préalables préparant une action classique de notre part (via notre corps). Enfin, comme nous le disions plus haut, il peut y avoir mise à disposition en parallèle, dans le monde, de l'action en mode classique et de l'action en mode numérique (ainsi, je peux toujours constituer une affiche en collant sur un support des morceaux de texte et d'image, même si, de nos jours, on fera beaucoup plus la chose avec un logiciel dit paginateur).

C'est un exercice de discernement vis-à-vis de la vie sociale en proie à la révolution informationnelle de distinguer les deux aspects : l'aspect tra-

duction vers le format numérique – qui correspond, mathématiquement, à le "reconstruction" de la classe de données comme une classe "constructive" d'objets, justiciable d'une définition récursive – et l'aspect définition d'un traitement de l'ordre de calcul, qui passe par la programmation d'une fonction. Dans certains cas, ce qui est génial est la traduction (comme, par exemple, dans le cas de la capture des sons et des images). Dans d'autres cas, ce qui est prodigieux est l'algorithme (si c'est un algorithme qui pilote une intervention chirurgicale par exemple). Nous sommes invités ainsi par la révolution informationnelle à une sorte de relecture mathématique du monde, mobilisant les mathématiques constructives et computationnelles.

La révolution informationnelle, par ailleurs, suscite une réflexion sur les limites de la mise en forme numérique du monde et de ses actions. Celles et ceux qui ont travaillé sur cette question me semblent arriver toujours aux mêmes résultats : la limite serait, en somme, triple. Ce que le régime du computationnel ne transcrit ou ne transpose pas, ce sont en effet le *continu*, le *corps* et l'*herméneutique*. Quelques mots de précision sont ici nécessaires.

Le continu. Les machines calculantes sont installées à demeure dans le fini et le discret. Aller vers le dénombrable suffit à révéler leur incompétence : une fonction quelconque n'est correctement réalisée par un programme que sur une distance finie. À vrai dire, on ne parvient à entrer en la machine les valeurs auxquelles appliquer une fonction, pour commencer, que tant qu'elles restent de complexité limitée. Pour prendre un exemple évocateur, lorsque vous produisez une suite $(u_n)_{n\in \mathrm{N}}$ telle que $u_{n+1}=f(u_n)$ en itérant indéfiniment l'action de f, au bout d'un certain temps – dans les cas convenables – votre machine affiche la limite théorique l satisfaisant $f(l)=l$, avec le degré d'approximation dont elle est capable, au lieu de persister à calculer des valeurs distinctes de l manifestant la convergence sur le mode asymptotique (en supposant que c'est ce que prévoit et décrit l'analyse réelle). De même, le théorème des valeurs intermédiaires est faux pour les valeurs effectivement connues ou atteintes par l'ordinateur (comme il l'est dans l'ensemble des rationnels **Q**, à vrai dire).

Nous constatons donc une incapacité foncière à exhiber, manifester le continu, à s'égaler à lui en aucune façon. Ce qui n'empêche pas de concevoir des programmes qui "vérifient" ou attestent les lois générales concernant le continu que la mathématique est capable de prouver. De manière liée, les calculs des ordinateurs peuvent servir de mode d'exploration pour mettre en lumière des comportements tendanciels des nombres décimaux, dont on voudrait établir qu'ils expriment des propriétés universelles nécessaires des nombres réels.

La difficulté philosophique s'associant à ce qui vient d'être dit est que l'on peut émettre deux objections au verdict de l'incompétence des ordinateurs au sujet du continu.

La première consiste à remarquer que le discours mathématique lui-même est débordé ou dépassé par le continu : il ne sait pas, par exemple, nommer chaque nombre réel. Tout le savoir de ce que nous visons comme l'ensemble transdénombrable des nombres réels passe par une expression théorique formelle et finie tout à fait de la même espèce que ce que l'ordinateur manipule et tolère. La différence, en fin de compte, réside dans l'effet intentionnel de la visée du continu transdénombrable : un effet que nous attribuons à notre discours mathématique, mais pour lequel nous ne trouvons pas d'équivalent du côté des machines.

La seconde consiste à faire valoir que, néanmoins, les simulations finies du continu dont les ordinateurs sont capables sont plus que satisfaisantes pour la grossièreté de nos perceptions : une image définie avec un nombre de pixels assez grand nous paraît d'une richesse "continue" bien autant que le réel lui-même (nous ne gardons avec nos organes perceptifs de l'une et de l'autre qu'une caricature finie).

Le débat peut ainsi se poursuivre, à trois niveaux en quelque sorte : débat sur la fictionnalité du continu mathématique, débat sur l'effectivité du continu physique, débat sur un éventuel continu cognitif. La limite du computationnel que serait le continu demeure incertaine et contestable.

L'herméneutique. L'argument, cette fois, est celui que Hubert Dreyfus a développé dans son célèbre *What Computers Can't Do*, une évaluation critique du projet de l'intelligence artificielle qui a été rééditée plusieurs fois[11]. Dreyfus constate d'abord que le projet de l'intelligence artificielle consiste à reconstruire tout comportement intelligent humain comme application de règles. Il observe ensuite que, dans les faits, lorsqu'on tente ce genre de transcription, on bute sur une difficulté que l'on peut identifier comme le besoin de méta-règles. Il y a bien, en effet, impliquées dans le comportement intelligent en cause, des règles que l'on suit, seulement la difficulté est que, en général, l'agent intelligent n'est pas supposé mettre en acte une telle règle dans tous les cas, aveuglément et sans égard à la situation. Il ne le fait que s'il se trouve dans des circonstances pertinentes pour l'application de la règle en cause. Il semble donc que le logiciel que l'on essaie d'écrire et qui récupèrerait notre intelligence usuelle

[11] Cf. H. Dreyfus, *Intelligence Artificielle, Mythes et Limites*, trad. franç. R.M. Vassallo-Villano, Flammarion, Paris 1984 ; édition américaine *What Computers Can't Do* (1972, 1979, 1992).

devrait disposer, en plus des règles dont l'application réalise la tâche considérée, de méta-règles stipulant dans quel cas "lancer" telle ou telle règle. Ce qui donne le sentiment d'être l'amorce d'une régression à l'infini.

Ce qui manque à la programmation d'une fonction récursive ou récursive primitive, en l'espèce, c'est une sorte de faculté d'évaluation holiste du contexte. Les paramètres à prendre en compte dans une telle évaluation ne se laissant pas confiner dans un registre déterminé : plus ou moins tout trait de l'environnement, mais aussi tout trait de la psychologie embarquée et tout trait de la normativité sociale encadrant celle-ci sont candidats à compter.

Dreyfus observe que dans le cas de l'agent humain, le problème est nativement résolu par la façon dont l'existence habite son monde, nage en lui. Il relie cette conception à la notion phénoménologique d'*être-au-monde*, qu'il trouve avantage à considérer dans sa version merleau-pontienne, où la prise sur le monde en situation est originairement le fait du corps.

Mais, comme il a été reconnu par nombre d'esprits réfléchissant à ces matières, et comme cela se dessine à vrai dire dès les formulations heideggériennes, le jeu de l'être-au-monde est foncièrement interprétatif. Savoir quelle règle appliquer revient à interpréter la situation : l'être-au-monde corporel, le *Dasein* incarné, produit toujours une telle interprétation par sa façon même d'en user avec son environnement, de naviguer en lui. Ce qui manque aux ordinateurs, c'est cet "en-vue-de-soi" des organismes humains qui s'exprime comme projection interprétative de leur monde, socle quasi-naturaliste annonçant les subtilités de la lecture des textes (où l'on reconnaît à l'œuvre, également, les fonctions de projection et d'anticipation mises en lumière par la tradition herméneutique, de Schleiermacher à Gadamer et Ricœur).

La limite du calcul comme forme et modèle de la pensée, et de l'objectivité constructive symbolique comme forme et modèle de la donnée, est donc double : l'objet peut prétendre échapper au nouveau régime informationnel en tant qu'objet relevant du continu et excédant à ce titre l'objectivité constructive, la pensée peut prétendre échapper au régime computationnel en tant que projection interprétative en situation excédant le mode "gouverné par des règles".

À quoi il faut ajouter, nous l'avons annoncé, une troisième limite : celle du corps, qui ne se confond pas avec les deux premières mais les confirme sans doute.

Le corps. L'informatique est installée dans le champ de l'idéalité symbolique, donnant lieu à la procédure récursive. Les contenus des mémoires

sont identifiés à des 0 ou des 1, valeurs fonctionnant comme des invariants au-dessus de leurs innombrables instanciations, dans une même mémoire, et d'une implantation de la machine théorique idéale à l'autre. Chaque programme auquel on donne le loisir de s'exécuter s'identifie à un texte, lui-même susceptible d'un nombre illimité d'instanciations, en machine et hors machine. Tout fragment de donnée ou tout morceau de procédure est répétable à l'envi, sans que son identité ne soit altérée par ces répétitions. Par conséquent, lorsque des données à l'origine non numériques sont portées au format numérique, ou lorsque des actions non computationnelles à l'origine sont traduites en computations numériques, c'est à chaque fois un élément du monde sensible et corporel ou un segment de devenir mondain qui sont transportés ou transposés dans le champ idéal du computationnel.

Interprété du côté de la “vie de l'esprit”, c'est-à-dire en comprenant la révolution informationnelle comme promouvant une version inédite du *Mind* – entité autrefois découverte et célébrée par les modernes – ce passage à l'idéalité via le computationnel signifie avant tout l'oubli du corps : de ce pré-sujet, sous-sol de la vie humaine, ancêtre et support de tout esprit, depuis lequel se détermine un rapport au monde dans la perspective duquel le monde est chair. C'est la philosophie de Merleau-Ponty qui dépeint au mieux ce “reste” – laissé de côté par la révolution informationnelle – qu'est le corps. Au gré de cette philosophie, d'ailleurs, l'oubli du corps tend à coïncider avec le manquement par rapport au mode herméneutique de la pensée.

4. *Les mathématiques et la vérité*

Si les mathématiques sont ainsi capables de faire problème pour la rationalité, pour l'épistèmè et sa philosophie – dont elles apparaissent pourtant, le plus souvent, comme le meilleur champion – c'est peut-être en raison de leur positionnement exceptionnel par rapport à l'enjeu suprême de toute connaissance, la vérité.

Pour le montrer, je vais dévoiler quelque peu ma pensée la plus personnelle, en faisant état de ce que j'appelle *ethanalyse de la vérité* (et dont on trouvera l'exposition pleine dans *Partages du sens*)[12].

[12] Cf. J.-M. Salanskis, *Partages du sens*, Presses Universitaires de Paris Nanterre, Nanterre 2014, pp. 109-147.

L'idée est que l'affaire de la vérité se laisse comprendre à la lumière de notre adhésion à une tradition séculaire : nous nous efforçons, depuis fort longtemps, de satisfaire à l'enjeu ou l'appel de la vérité, tel qu'il résonne auprès de nous à partir du mot lui-même. Un mot exceptionnel en ce que, pour une part, sa valeur n'est pas extensionnelle : le mot vaut comme ce que j'appelle un *sollicitant*, c'est-à-dire qu'il exprime un appel, nous l'entendons comme nous demandant de rejoindre la hauteur de son enjeu. Nous partageons le sens de la vérité dans ce que j'appelle un *ethos* : un ensemble comportemental admettant les exigences de la vérité comme ce par rapport à quoi l'on se mesure (le *nous* nommé au début de la phrase est plus précisément celui des *adeptes* de l'ethos, celles et ceux qui sont sensibles à l'appel). L'ensemble comportemental en cause, bien qu'il puisse être constitué d'actes déficients par rapport à l'enjeu de la vérité, persiste à se situer par rapport à lui : au sein de l'*ethos* de la vérité on juge les comportements par rapport à l'enjeu de vérité.

Clairement, ce qui précède présuppose déjà silencieusement que l'appel de la vérité a été converti en une liste explicite d'exigences : sinon, se mesurer à l'enjeu de la vérité ne signifierait rien de suffisamment déterminé. L'ethanalyse est l'enquête philosophique qui, pour chaque "appel" véhiculé par un sollicitant, travaille à exhiber la liste des prescriptions déterminant implicitement entre nous la fidélité à l'enjeu : une liste que j'appelle la *sémance* de l'ethos. Un ethos, à chaque fois, possède aussi une dimension historique : ses adeptes transmettent l'enjeu aux générations ultérieures, font en sorte que l'on continue de se rattacher aux exigences de la sémance, que l'on persiste à partager le sens en cause.

Néanmoins, par essence, pour autant qu'ils dépendent de notre constance, de notre fidélité, les ethos sont des traditions contingentes. Et il en va ainsi de l'ethos de la vérité, même s'il a traversé les siècles et si nous sommes nombreux à "miser sur lui". Même s'il semble s'être acquis une sorte de place enviable, de prestige dans notre société. Cela n'interdit pas de, s'inquiéter de possibilités nouvelles de l'oublier qui se font jour.

Dans mon ouvrage de 2014, j'expose six prescriptions constituant la sémance de la vérité :

Deux prescriptions touchant le registre "ontologique" :

1) Regarde l'être comme étranger.
2) Envisage l'instance de la connaissance comme hors-être.

Ces deux prescriptions correspondent à la garde de la définition adéquationniste de la vérité. Nous devons regarder le réel comme non docile a

priori à quoi que ce soit que nos humeurs ou intérêts puissent requérir. Et la connaissance est une performance qui n’inclut pas le sujet, qui le laisse dans un mystérieux écart vis-à-vis de toute réalité, au moment où elle détermine cette dernière. Si je tente une connaissance du rapport du savoir à l’être, je suscite une nouvelle instance hors être à l’arrière-plan de cette tentative.

Deux prescriptions touchant le registre “épistémologique” :

3) Accueille l’être étranger, comme un maître d’étude.
4) Pense et élabore une identité de la prestation de connaissance vers l’être accueilli.

Ces deux prescriptions élaborent l’idée d’une expérience et de l’émission d’un jugement vrai. Pour que nous puissions dire vrai, il faut que nous ayons accueilli l’être étranger. Mais cela ne saurait avoir le sens d’un événement du réel, dont nous aurions la science avant la connaissance : plutôt, l’étrangeté s’impose à nous dans un accueil suivant lequel elle entre en nous comme information, exactement comme le maître d’études fait venir en nous des pensées que nous n’avions pas déjà. Reste alors à concevoir un mode et une règle de l’identité de ce que nous disons dans le langage et de ce qui est accueilli (dont la forme doit se prêter à une telle identité).

Deux prescriptions touchant le registre “pragmatique” :

5) Assume l’engagement dans la vérité de ton discours.
6) Ecoute le niveau décontextualisé de tout dit.

Ces deux prescriptions élaborent l’idée que notre pratique du langage “repose dans la vérité” (ou du moins, est appelée à “reposer dans la vérité”). Ce que nous déclarons, nous le déclarons forcément comme assertion vraie auprès du réel. Et ce que nous énonçons, s’il dépend d’un contexte, renvoie à une énonciation de cette dépendance elle-même qui vaut hors tout contexte (de même que, lorsque je prouve B sous l’hypothèse A, j’ai prouvé $A \rightarrow B$ sans hypothèse).

La question que l’on peut poser, à ce stade, c’est si les mathématiques soulèvent une difficulté à ce niveau absolument général de la structure normative de la vérité. S’il en va bien ainsi, cela procure, peut-être, un éclairage sur le statut paradoxal et exceptionnel des mathématiques par rapport à l’intention d’une philosophie des sciences.

Pour résumer mes conclusions, il apparaît à l’examen que les mathématiques fournissent un cas dégénéré d’observance de l’injonction 1 de

la sémance. À certains égards pour cette raison même, elles offrent au contraire une observance "superlative" de l'injonction 4.

En effet, pour nous concentrer d'abord sur le point de l'injonction 1, comment la mathématique peut-elle se concevoir comme connaissance d'un "être étranger" ?

On accordera au "tournant linguistique" du vingtième siècle que l'instance du connaître, pour nous, ne peut être que langagière. Il est apparu assez clairement, il me semble, et beaucoup de philosophes l'ont dit de beaucoup de façons, que l'instance du connaître n'est pas un sujet psychologique, mais correspond plutôt à un ensemble de procédures et à un horizon d'expression qui sont ceux du langage. De plus, en situant résolument toute mathématique dans le contexte d'un langage formel, l'approche contemporaine a pour ainsi dire souligné et officialisé ce point dans le cas des mathématiques. Donc, ce par rapport à quoi l'être "à connaître" doit se manifester comme étranger, c'est la réalité linguistique, c'est-à-dire une réalité de type idéal, de part en part : les unités constituantes des langages sont des types, et les règles des langages sont des règles s'appliquant à des répétitions de cas et des règles dont la forme se répète dans chaque cas. En telle sorte que, c'est une partie de ce que la pensée contemporaine nous a enseigné et fait comprendre, la réalité linguistique est tout entière prise dans le jeu idéal du type, de l'occurrence et de la répétition ne compromettant pas l'invariant qu'est le type. Un théorème s'identifie par suite à une forme d'assemblage idéale d'unités idéales, dominant une diversité illimitée d'occurrences. En ce sens, la connaissance mathématique "incarne" de façon parfaitement satisfaisante le "hors être" de l'instance du connaître dont parle l'injonction 2. Avant tout formalisme, à vrai dire, il n'y a déjà pas de place dans notre ontologie pour l'invariant du mot *maison*, seulement "présenté" par chacune de ses occurrences, avec laquelle il ne s'identifie pas.

Intéressons nous alors d'abord à l'objectivité la plus basique et la plus inaliénable dont traite la mathématique[13] : l'objectivité qu'on peut appeler *constructive*, celle que Brouwer avait mise en avant (et celle sous laquelle se rangent les données que nous fournissons à nos ordinateurs). En effet, cette objectivité, qui est celle des objets qui se laissent construire conformément aux clauses d'une définition récursive, se trouve être "déjà", ou "originairement" si l'on préfère, l'objectivité des expressions formelles.

[13] Pour tout ce qui suit, cf. J.-M. Salanskis, *Philosophie des mathématiques*, Vrin, Paris 2008, pp. 35-108.

Les termes, formules et preuves d'un langage formel relèvent d'une définition récursive comme les entiers naturels dans la perspective constructive. Par suite, on peut redouter que la connaissance mathématique, comme connaissance constructive de l'objet constructif, soit toujours seulement redoublement spéculaire : réflexion par le langage de ses propres structures. De fait, si l'on considère la preuve formelle de 7+5=12 dans le système PA (l'arithmétique formelle de Peano), on observe qu'elle répète, dans l'économie des lignes successives dont elle se compose, la synthèse a priori de 12 à partir de 7 et 5 autrefois décrite par Kant, qui est la même chose que l'élaboration de 7+5 comme indication de construction conduisant au même construit que 12 (la preuve, simplement, utilise cinq fois l'axiome $x+Sy=S(x+y)$). La connaissance, en l'espèce, se montre, comme on pouvait le redouter, spéculaire. Jusque là, il peut sembler que l'injonction de l'étrangeté de l'être à connaître soit purement et simplement ignorée.

Néanmoins, au moyen de l'arithmétique formelle – intuitionniste ou non – sont développés des outils et obtenus des résultats qui affrontent les configurations arbitraires concernant des constructions de longueur ou de complexité illimitée. Même la connaissance constructive de l'objectivité constructive s'avère ainsi tournée vers la complexité, mode de l'“immaîtrisable jusqu'à nouvel ordre” dans le champ de l'objectivité constructive. En telle sorte que, déjà au plan de ce sous-sol de la connaissance mathématique, l'“être” à connaître – qui, d'abord, n'est autre que l'idéalité des configurations arithmético-linguistiques – prend finalement une figure d'étrangeté : soit l'étrangeté minimale liée au fait qu'il est cité et mis en scène comme objectivité justement (et pas seulement vécu sur le mode opérationnel comme il en va ordinairement avec les entités linguistiques), soit l'étrangeté supplémentaire et décisive liée à l'horizon de la complexité. Le nom 3 – répétable à l'envi – se distingue de l'insaisissable qu'est l'entier naturel successeur du successeur du successeur de 0 “lui-même”, et la preuve que la somme des n premiers entiers vaut $n(n+1)/2$ embrasse la complexité (à une profondeur ou échelle arbitraire). Ce dont traite la mathématique constructive, c'est ainsi du champ d'objets enveloppé dans notre opérativité langagière originaire, “rehaussé” ou transfiguré dans une posture d'être étranger par son installation au pôle référentiel ou intentionnel, et par l'horizon de la complexité : par un regard cherchant en lui l'asymptotique de ce qui se tisse en lui.

Mais nous devons évoquer aussi le second versant de la connaissance mathématique, ce que j'appelle le versant de l'objectivité corrélative. Cette fois, l'objet dont traite la mathématique est pris comme membre de

multiplicités convoquées sur un mode pour ainsi dire "métaphysique" par une stipulation axiomatique. On peut en effet aussi, c'est en substance ce à quoi Hilbert nous a encouragés en définissant la méthodologie formaliste, introduire "d'un seul coup" la classe des objets dont on veut traiter par la stipulation dans une liste d'axiomes des propriétés que l'on souhaite voir satisfaites collectivement par eux (les axiomes, en règle générale, comportent une quantification implicite ou explicite mobilisant l'universel sur le mode du $\forall$ ou du $\exists$). Et ce que l'on appelle connaître de tels objets coïncide alors avec la déduction de théorèmes dans la théorie logique spécifiée par la base axiomatique adoptée.

Un tel mode du connaître identifie bien, de nouveau, l'instance du connaître à l'idéalité d'un langage formel : les interventions exprimant la connaissance sont des textes listant des formules du langage considérée, s'achevant dans une formule prenant le nom de *théorème*, distinguée par là comme disant le vrai. Mais l'être à connaître revêt-il le statut d'être étranger ?

Il le fait en un triple sens : 1) d'un côté, le rapport des mathématiciens à l'univers de corrélation s'associant à la stipulation axiomatique (typiquement, l'univers des ensembles suscité par la stipulation de la théorie ZFC) est un rapport phénoménologique de projection, prenant même la valeur d'un être-au-monde pour eux ; 2) d'un second côté, l'univers stipulé est stipulé devoir inclure en son sein le déploiement de toute objectivité constructive, et la "vérité" associée du prouvable est supposée inclure comme cas particulier la (confirmation de la) vérité constructive. Donc l'univers auquel s'adresse la connaissance selon le schéma de l'objectivité corrélative enveloppe la complexité constructive, mode de l'étrangeté de l'être à l'étage inférieur ; 3) d'un troisième côté, la connaissance mathématique de l'objectivité corrélative entre dans le "jeu de l'infini".

Pour développer le point 1) : les multiplicités projetées depuis des stipulations sont analysées dans leur variation possible depuis une théorie interne à ZFC qui est la théorie des modèles. Dans cette guise, et notamment à la suite des travaux de Kripke, un univers de corrélation, dans la version décalée qu'en donne la notion de modèle, devient l'expression mathématique de l'idée d'un "monde possible". La pensée de l'objectivité corrélative met en scène l'être à connaître dans son étrangeté modale (il pourrait toujours être autre). D'autre part, non seulement le mathématicien "projette" en quelque sorte l'univers de corrélation à partir de la stipulation, lui donnant un statut ressemblant à celui du noématique chez Husserl (juste, c'est un noématique collectif enveloppant l'objectivité

constructive et conservant la vérité constructive), mais encore, les axiomes ouvrent la possibilité de “gestes” dans l’espace théorique qui sont vécus comme des gestes dans l’univers stipulé. Typiquement, comme je l’explique dans *Philosophie des mathématiques*, un groupe est un domaine dans lequel je peux simplifier à droite ou à gauche[14]. J’associe donc un *faire* aux univers de corrélation, ce qui perfectionne l’illusion qui leur attache une étrangeté de monde.

Pour ce qui concerne le point 2) : en raison de l’axiome de l’infini (axiome qui pose, parmi la faune de l’objectivité corrélative ensembliste, un ensemble infini), on est assuré a priori que tout déploiement d’objectivité constructive trouvera sa place dans l’univers des ensembles. À vrai dire, tel est en substance le sens théorique donné par la stipulation au concept d’infini : un ensemble infini est un ensemble au sein duquel se laisse inclure une série récursive d’objets. Par conséquent, le “pseudo-être” intentionnel que s’est donné la stipulation ensembliste – lui conférant une étrangeté intentionnelle/modale de monde – enveloppe par principe l’horizon d’étrangeté de la complexité, évoqué à l’instant.

Pour considérer enfin le point 3) : la stipulation majeure de la mathématique contemporaine, celle de la théorie ZFC, nous avons déjà dû le dire pour expliquer qu’elle a affaire à la complexité illimitée, nous engage dans le jeu de l’infini. Ce qui veut dire non seulement qu’elle pose l’infini actuel, mais qu’en elle se développe (se retraduit) la théorie cantorienne du transfini. De la sorte, au déploiement de l’univers ensembliste s’associe aussi le jeu d’une transcendance : transcendance vis-à-vis de toute complexité constructive d’abord, mais aussi transcendance à l’égard de soi-même, empêchant que l’infini puisse jamais se stabiliser à un niveau ultime quelconque. L’étrangeté du pseudo-être dévisagé, prend ainsi une figure radicale, celle de l’infini, dont l’énigme défie la pensée depuis toujours.

De fait, le débat épistémologique du XX^{e} siècle témoigne de ce que, avec l’infini, la mathématique est entrée dans un régime portant à l’extrême le “fictionnel” que l’on pouvait depuis toujours lui imputer. Du point de vue de l’ethanalyse de la vérité, cela dit, un tel “excès” n’empêche pas, loin s’en faut, que l’être étranger soit pris comme maître d’étude, et que son étrangeté prenne précisément ce sens pour le connaître qui la poursuit : les divers axiomes supposés exprimer un mode ou une loi du transfini ensembliste (comme les axiomes de grands cardinaux, dont l’axiome de l’infini est le premier), fixent dans le langage et dans la théorie

[14] Cf. *Philosophie des mathématiques*, cit., pp. 85-86.

une notion de transcendance "entendue" comme caractéristique de l'infini se proposant dans son étrangeté.

En bilan, on voit que la "vérité mathématique" tout à la fois s'inscrit dans l'ethos de la vérité, et en est un cas dégénéré vis-à-vis de l'injonction 1, parce que l'"être mathématique" prend d'abord une figure interne et proche. Il est originairement, comme être de l'objectivité constructive, homogène à l'élément actif intime du connaître (le linguistique constructif), ce qui semble interdire l'étrangeté ; mais celle-ci se voit sauvée comme étrangeté de principe du référent et comme apportée par l'horizon de la complexité. D'une seconde manière, comme être de l'objectivité corrélative, l'être mathématique apparaît comme non étranger aussi, dans la mesure où il est stipulé, et ce qui est vrai à son égard paraît commandé en interne par la structure idéale du langage. Mais, de nouveau, il récupère tout de même l'étrangeté comme projection intentionnelle, comme variation modale de monde, ou comme infini.

Parallèlement, nous comprenons aussi comment la vérité mathématique peut, dans cette condition paradoxale, être exemplaire. L'identité voulue par la clause 4 est parfaite dans le cas fondamental de la vérité spéculaire de la mathématique constructive ; et l'identité de la vérité est également sans faille dans le cas de l'objectivité corrélative, parce que l'être et sa configuration sont institués comme répondant sans écart et sans différence au dit et à sa forme. Par dessus le marché, la prouvabilité dans la théorie mettant en scène l'objectivité corrélative est supposée "conserver", c'est-à-dire reprendre et retrouver dans son régime autre, la vérité constructive spéculaire fondamentale : c'est ce que Hilbert a formulé comme l'exigence dont nous devions garantir a priori le respect, dans son fameux "programme". Nous nous autorisons donc à vivre la vérité mathématique comme une vérité disant une identité aussi rigide et absolue que l'identité spéculaire du cas constructif fondamental.

Cela explique le fait, souvent évoqué et commenté, qu'il nous est plus facile de remettre en question nos données perceptives que nos certitudes mathématiques : celles-ci ont un statut de super-vérités trans-empiriques souligné par des générations de philosophes, depuis Platon jusqu'à David Armstrong en passant par Kant.

Le tableau est donc bien celui que nous annoncions : la discipline qui ne satisfait que de manière dégénérée à l'injonction 1 de la sémance de la vérité est en même temps celle qui observe de la façon la plus généreuse et superlative son injonction 4.

5. *Conclusion*

Toute philosophie n’est pas attentive à ce statut paradoxal, statut de supplémentarité et d’exception, de la mathématique. La mathématique est pourtant sans doute, comme l’a dit Platon dès l’origine de la philosophie, la propédeutique par excellence de la philosophie. Mais elle l’est tout autant parce qu’elle nous enseigne l’inextricable, l’excès, l’étrangeté, que parce qu’elle nous détourne du sens commun et nous oriente vers l’idéalité. Les problèmes qu’elle soulève pour une philosophie cherchant à penser la science, ou le monde avec la science, sont également une manifestation supplémentaire de l’école d’humilité qu’est la mathématique. Pour qui a eu assez de bonnes dispositions envers les mathématiques pour découvrir quelques avenues de leur splendeur – coïncidant avec leur infinie, leur labyrinthique difficulté – il est bien difficile de ne pas retenir la leçon de la fragilité et de la modestie de nos pouvoirs intellectuels.

Qu’il me soit permis de noter, pour conclure, qu’à côté de l’orientation historique qui lui est très souvent reconnue, la philosophie française contemporaine des sciences, telle qu’elle a été formulée par plusieurs générations, de Brunschvicg à Granger et Vuillemin en passant par Cavaillès, Lautman et Desanti, a montré il me semble une sensibilité aux difficultés philosophiques liée aux mathématiques : on pourrait y voir sa signature, l’expression de son feeling propre, aussi légitimement que lorsqu’on les localise dans l’esprit historique. Et cette sensibilité, à mon avis, était liée à une reconnaissance de la grandeur déconcertante de la pensée mathématique.

En tout état de cause, on est en droit de demander à toute philosophie des sciences, il me semble, de ne pas passer à côté du “problème des mathématiques”.

English title: The issue of mathematics.

Abstract

This paper sustains that mathematics is a deep source of difficulties for philosophy of science in general. First, it is not easy for philosophy, while recognizing what it owes to mathematics, to locate itself with respect to it. Second, the main debate of philosophy of science – which opposes realism and something like projective rationalism – is governed by how we under-

stand the role of mathematics. Third, we have to refer to mathematics in order to build a correct picture of informational revolution, even if mathematics could count as a reason for resisting that revolution.

In its last part, the paper explains all these difficulties as rooted in the exceptional way mathematics satisfies the demands of truth.

Keywords: mathematics; philosophy; Kant; Frege; metaphysics; computers; continuum; hermeneutics; body; constructivism.

Jean-Michel Salanskis
Université Paris-Nanterre
jmsalanskis@gmail.com

Edizioni ETS
Palazzo Roncioni - Lungarno Mediceo, 16, I-56127 Pisa
info@edizioniets.com - www.edizioniets.com
Finito di stampare nel mese di aprile 2019

TEORIA

T

Rivista di filosofia
fondata da Vittorio Sainati

Ultimi fascicoli apparsi della Terza serie di «Teoria»:

XXXVIII/2018/2 (Terza serie XIII/2)
Virtue Ethics / Etica delle virtù

XXXVIII/2018/1 (Terza serie XIII/1)
Back to Ancient Questions?
Tornare alle domande degli Antichi?

XXXVII/2017/2 (Terza serie XII/2)
Etica, diritto e scienza cognitiva / Ethics, Law, and Cognitive Science

XXXVII/2017/1 (Terza serie XII/1)
Linguaggio e verità / Language and Truth

XXXVI/2016/2 (Terza serie XI/2)
Etiche applicate / Applied Ethics

XXXVI/2016/1 (Terza serie XI/1)
New Perspectives on Dialogue / Nuove prospettive sul dialogo

XXXV/2015/2 (Terza serie X/2)
Relazione e intersoggettività: prospettive filosofiche
Relación e intersubjetividad: perspectivas filosóficas
Relation and Intersubjectivity: Philosophical Perspectives

XXXV/2015/1 (Terza serie X/1)
Soggettività e assoluto / Subjectivity and the absolute

XXXIV/2014/2 (Terza serie IX/2)
«Ripensare la 'natura' – Rethinking 'Nature'
2. Authors and Problems/Figure e problemi»

XXXIV/2014/1 (Terza serie IX/1)
«Ripensare la 'natura' – Rethinking 'Nature'
1. Questioni aperte/Burning Issues»

XXXIII/2013/2 (Terza serie VIII/2)
«Hope and the human condition – Speranza e condizione umana»

XXXIII/2013/1 (Terza serie VIII/1)
«Hegel. *Scienza della logica*»

Questo fascicolo di «Teoria» vuole riflettere sul tema della fiducia a partire dal suo significato etimologico e tenendo conto dei molteplici ambiti in cui tale atteggiamento viene a giocare il suo ruolo nelle relazioni interumane. Si tratta del primo di due volumi di «Teoria» dedicati all'argomento. In questo fascicolo la fiducia viene approfondita in una prospettiva teorica e con un approccio multidisciplinare. Il prossimo si concentrerà su alcuni autori della storia del pensiero nei cui lavori tale questione è stata esplicitamente affrontata.

This issue of «Teoria» is a reflection on the theme of trust, starting from its etymological meaning and considering the multiple areas in which trust plays a role in inter-human relations. This is the first of two volumes of "Theory" devoted to the theme. In this issue the theme of trust is looked at more closely from a theoretical perspective and through a multidisciplinary approach. The next issue will focus on some authors of the history of thought in whose work this question has been explicitly addressed.

ISSN 1122-1259

ISBN-13: 978-8846755179