Marco Brigaglia

Obedience and control: insights from the cognitive sciences

English version of the chapter *Obbedienza e controllo: spunti dalle scienze della mente*, in Brigaglia M., Pino. G., Vallini A., *Obbedienza e responsabilità*, Roma Tre Press 2026, pp. 33–54.

ABSTRACT

In this chapter, I focus on the importance of psychological inquiry for the problem of responsible obedience. I outline a possible, desirable development of the ABIDE project (of which this book is an output), highlighting its main methodological challenges. I present the reader with a critical reconstruction of Milgram's experiments and a series of recent studies conducted by Caspar, Haggard, and colleagues based on an innovative experimental paradigm, which have given new vigour and a fresh perspective to Milgram's hypotheses on the mechanisms of obedience.

KEYWORDS

Obedience and control; Milgram; Psychology of obedience; Responsible obedience.

Funded by the European Union - NextGenerationEU under the National Recovery and Resilience Plan (PNRR) – Mission 4 Education and research – Component 2 From research to business - Investment 1.1, Notice Prin 2022 indetto con DD N. 1409 del 14/9/2022, titled "Rule of Law and the Problem of Responsible Obedience (ABIDE)", proposal code P20229FK2F - CUP B53D23032560001.

Marco Brigaglia[1]

Obedience and voluntary control: insights from the cognitive sciences

SUMMARY: 1. Introduction: the problem of responsible-yet-disciplined obedience – 2. Proposals for future work – 3. Milgram on the 'alienation of control' – 4. The experimental paradigm of Caspar and colleagues: sense of agency and intentional binding – 5. Experiments and results – 6. Conclusions and some comments.

1. Introduction: the problem of disciplined-yet-responsible obedience

A central aspect of the ideal of 'responsible' obedience is the maintenance, by the subject to whom an order – or, more generally, an authoritative directive – of a margin of deliberative control over the process of selecting their own action, which allows them, first and foremost, to subject the content of the order to critical scrutiny, and which also allows them, in the event of a serious flaw, to resist the psychological pressure exerted by the order, choosing to disobey. The ideal of responsible obedience therefore requires the individual to *reflect* on the content of the order and its correctness, and to be able *to resist* the pressure of the order.

On the other hand, 'discipline' – in one of the meanings of the term – is the acquisition of a *habitus* of obedience, understood as a vast complex of attitudes and dispositions that include, amongst other things, a certain degree of 'automatisation' of obedience: the tendency to carry out orders received promptly, smoothly and immediately, thereby, to a certain extent, removing the order from one's own critical scrutiny. 'Discipline', in this sense, consists in the acquisition of the tendency *not* to *resist* the order and to carry it out quickly and *without reflection* – or at least, without reflecting too much[2].

---

[1] Full Professor of Philosophy of Law, University of Palermo.

[2] In Weber's terms, 'discipline' is the disposition, acquired through habit, towards 'prompt', 'automatic' and 'stereotypical' obedience (M. WEBER, *Economy and Society*, Mohr, Tübingen 1922, p. 28). Further on, the construction of a 'rationalised' discipline, characteristic of large organisations in the modern age, is defined by Weber in terms of the exact and scrupulous execution of orders, in which any internal criticism of the order is suspended and the agent concentrates exclusively on the objective of completing the task (ibid., 642). Even if not necessarily in such a stringent

Between these two requirements – the requirement for 'accountability' in obedience and the requirement for discipline as a partial automation of obedience – there is an evident tension: a specific aspect of the more general tension that afflicts any organisational structure, between, on the one hand, the need for flexibility and accuracy in decision-making, and, on the other, the need for stability and speed.

The tension between responsible obedience and discipline is particularly significant in a political and legal context inspired by the principles of the *rule of law* and constitutionalism; and, although it affects the functioning of any administration, whether public or private[3], it takes on a dramatic dimension within the military system (and, more generally, across the entire 'security sector'). In this context, there is a particularly strong need for discipline – and for *the cultivation* of discipline through specific training and organisational methods – to ensure effective coordination of action. But equally strong is the need to ensure that discipline does not transform the armed forces, and the immense power they wield, into an instrument of illegality or even subversion. The ideal of responsible obedience thus becomes, in this context, part of a political project to 'constitutionalise' the armed forces (and, more generally, the entire security sector), including through widespread scrutiny, exercised by subordinates, of the legitimacy and constitutionality of orders.

This political project presupposes that the requirements of discipline and of responsible obedience (and, in particular, of a responsibility informed by the principles of the *rule of law* and constitutionalism), despite their evident tension, are not ultimately incompatible, but can find a satisfactory point of compromise. Herein lies a crucial component of what, in the ABIDE project, we have termed the *'problem* of responsible obedience': whether this point of compromise exists, what it might be, and how to achieve it. That is to say, whether and how it is possible to reconcile the need to *automate* obedience, at least in part, with the need to maintain a space for *deliberative control* over it.

2. Proposals for future work

To address the problem of disciplined-yet-responsible obedience seriously, it should be clear that it is not enough to proclaim the ideal in vague terms and prescribe it normatively to subordinates. Instead, a much more systematic and complex effort is required, comprising at least three steps:

---

form, discipline nonetheless requires some restriction of the scope for deliberation on the part of the recipient of the orders.

[3] See, in this regard, the contribution by V. MILITELLO in this volume.

(1) Specify much more precisely what model of obedience is being sought; which, among the many theoretically possible options, is the balance being sought between discipline and responsibility, and between automaticity and control. Furthermore, to be acceptable, the outlined model of obedience should not merely be abstractly satisfactory on the basis of our normative ideals, but should also be *realistic* – it should be possible, and not excessively difficult, for the decision-making processes of a human being of normal cognitive ability, in the relevant institutional contexts, to conform to the model (perhaps with appropriate education or training).

(2) Plan and establish organisational contexts and methods of education and training that promote the acquisition of the form of obedience specified by the chosen model.

(3) Design and implement, on the basis of the chosen model of obedience, a comprehensive and coherent reform of its legal framework, with particular attention to criminal liability.

The work carried out by the ABIDE project has, to date, been predominantly exploratory, reconstructing the way in which the issue of responsible obedience is currently addressed in various sectors of the legal system and within national and supranational institutional contexts, identifying limitations, grey areas and difficulties. Any continuation of the project should decisively move on to *the pars construens*, beginning to engage directly in the three steps outlined above[4] .

The first of these steps – defining the model of obedience being pursued – requires a great deal of theoretical groundwork beforehand. The difference between 'responsible' obedience on the one hand, and the type of obedience dismissed as 'automatic' ('mechanical', 'blind', 'passive', 'lifeless') on the other, is not a clear-cut distinction, but a complex and multidimensional continuum: control can take on different degrees and intensities, and automaticity can affect, to varying degrees, different aspects of the decision-making process. Purely automatic obedience, in which one person is acted upon by another in a purely passive manner, without any intervention of their own capacity for voluntary control, understanding and reasoning, is an extreme pathological condition, far removed from the forms of obedience found in ordinary experience – although, perhaps, revealing certain basic mechanisms that predispose one to obedience and which, under normal conditions, are mediated and

[4] Insights along these lines are contained in the chapter by A. SPENA in this volume. Spena identifies two dominant models of obedience, which he calls the 'model of *blind* obedience' and the 'model of *judicious* obedience'; he shows how the regulation of obedience as articulated by domestic criminal law oscillates, with little consistency, between the first and second models; he suggests that neither model is normatively or psychologically adequate – the first being too heavily weighted towards automatism, and the second too heavily weighted towards control; and finally, he outlines an alternative model, the 'model of *tragic* obedience'.

modulated by higher-level capacities[5] . Generally, obedience does not exclude but requires some form of voluntary control – the agent understands the prescribed behavioural model, transforming it into a criterion for action that guides their behaviour consciously, and often through the exercise of attention and willpower. At the same time (and this is the case with the experience we tend to describe as 'passive'), obedience may be accompanied by a abdication of the exercise of a higher level of control: the relinquishment of deciding how to behave on the basis of one's own arbitrary choice, and, above all, the relinquishment of deciding on the basis of one's own judgement, of one's own assessment of the applicable reasons – that is, the relinquishment of the exercise of 'deliberative' control. But here too – we shall see an example of this with the subjects of the Milgram experiment – things are by no means simple. There are cases in which the abdication of deliberative control is, in turn, a deliberate choice, which does not seem to express a state of passivity, but rather a state of extreme (albeit paradoxical) mastery over oneself. Consider, as an extreme case, 'ascetic' obedience, accepted as part of a voluntary path of detachment from the self; or the absolute obedience shown by the Samurai to his *damiō*, which (in its ideal-typical, or perhaps ideological, representation) is pursued as the highest expression not of servitude, but of self-control. But, even beyond these extreme conditions, there are cases in which the abdication of deliberative control appears to be the result of a deliberation based on a normative system with which the subject fully identifies – the agent is convinced that the right thing to do is to carry out the order without questioning it. At other times, however, the subject who obeys seems trapped in a mechanism with which they do not identify, but from which they are unable to escape (this is what appears to happen to the subjects of Milgram's experiment). Moreover, the relinquishment or compromise of deliberative control that accompanies certain forms of obedience can in turn take very different forms: a tendency *not* to *think*, to take the received order as the criterion for action without even considering further reasons that might suggest disobedience, or a tendency to *shield* one's actual decision from the influence of reasons other than the order received[6] – and this tendency can be so strong as to make obedience, in a sense, 'automatic', though not necessarily 'blind' (the agent internally *evaluates* the merits of the order, but acts on the basis of the order and not of their evaluation of its merits). Furthermore, obedience can be 'enthusiastic', accompanied by a

[5] In catatonia (a pathological condition involving severe impairment of voluntary control over action), an extreme form of automatic obedience may occur: the patient passively conforms, without hesitation or resistance, to the instructions of others, even when these require the adoption of humiliating, unusual or disadvantageous behaviours. It is possible that, in this condition, the temporary impairment of the capacity for voluntary control frees up a tendency towards the automatic execution of orders which, under normal conditions, is filtered, modulated and inhibited by control processes at various levels.

[6] J. Raz, *Practical Reason and Norms*, 2nd ed., Oxford University Press, Oxford 1999, p. 184.

complete and creative adherence to the authority's aims. Or it may instead occur 'reluctantly', limited to the bare minimum and accompanied by feelings (and signs) of unease, resistance, and discomfort. The functioning of mechanisms of automatic obedience can also vary greatly. I am referring to mechanisms, whether innate or acquired through learning (especially *training*), which operate by stimulating, in response to the receipt of an order, the automatic (immediate, involuntary, non-deliberate) adoption of compliant behaviour. In certain cases, these mechanisms may be limited to *facilitating* obedience, leaving intact the subject's ability to exercise control and, if necessary, inhibit obedience. In other cases, their effect may be more pervasive, making the exercise of control particularly difficult. Furthermore, the deliberative control that the subject exercises over the order can take various forms and modes, combined in the most diverse ways: an 'explicit' mode, involving more or less extensive conscious and deliberate reflection, or an 'implicit' mode, based on sensations and intuitions.

In short, the field of obedience is extremely diverse, and it is necessary, first and foremost, to undertake a theoretical exercise in mapping the various forms of obedience, highlighting their most significant normative and organisational aspects. However, in order to contribute to the construction of a *realistic* model of responsible obedience, this mapping must be *psychologically plausible*. And to achieve a psychologically plausible mapping, one cannot limit oneself to investigating obedience whilst remaining within the confines of common-sense psychology (the set of schemas, models, concepts and metaphors with which we ordinarily represent mental activity) and the kind of introspective reflection and observation we carry out based on it – the kind of approach that characterises most philosophical and legal reflection on obedience. Instead, it is necessary to integrate into one's approach a close engagement with (neuro-)psychological research into the dynamics of obedience and decision-making processes in general, and this requires opening up to the level of description offered by the cognitive sciences, which is often far more accurate and, in any case, calibrated according to the possibility of empirical ('operationalised') verification.

Consider, for example, the difference – to which I have repeatedly referred in the preceding pages – between thought and action that take place under our conscious and voluntary control, and those that are triggered and proceed 'automatically', without, and even against, our efforts to control them. This is a distinction that forms part of common-sense psychology, and is therefore very familiar. But it is also a distinction which, over the last fifty years, has been the subject of intense investigation by the cognitive sciences, and this investigation has not only produced explanations of the mechanisms underlying automaticity and control, but has also, to a large extent, enriched, specified, complicated, and in part even transformed the very notions of automaticity and control. We must endeavour, as far as possible, to move from the common-sense use of these concepts – useful

for providing an initial framework for understanding the phenomena, yet irremediably metaphorical and imprecise – to a usage closer to the scientific one. Consider, once again, the various, entirely metaphorical, terms used to capture the opposite pole of responsible obedience: 'automatic' obedience (like an automaton, without judgement or understanding), 'mechanical' (like a machine, also devoid of judgement and understanding), 'blind' (which sees no reasons beyond the order), 'cadaverous' (like a soulless body), 'passive' (like a body acted upon by others). A close examination of models from the cognitive sciences allows us to transform these metaphors – illuminating, yet also vague and often misleading – into more precise and informative concepts.

Let me give another example. An optimal solution to the problem of disciplined-yet-responsible obedience would require, amongst other things, that individuals acquire the most accurate possible ability to distinguish, in a semi-automatic manner (quickly and with a low level of consciousness and attention, 'intuitively'), between 'normal' cases, in which orders or instructions show no signs of dubious legitimacy or other flaws, from 'problematic', suspicious cases (the threshold of 'objective manifest criminality' of an act may be acceptable as a condition for punishability, but it is too low in relation to the principle of the proper functioning of public administration), responding in normal cases with the immediate execution of the order or instruction, avoiding engaging in the search for and scrutiny of reasons against execution (in such cases there is no need for reflection: obedience may be 'automatic', although, to the extent that it is underpinned by appropriate discernment, it cannot be considered 'blind'), and responding instead, in problematic cases, with the provisional suspension of execution and the initiation of a process of reconsideration and critical scrutiny of the order[7] . Traditional philosophical reflection, centred on common-sense psychology, can play an important role in pointing to this phenomenon in a way that makes it recognisable (for example, through the idea of a non-deliberative 'responsiveness to reasons', a capacity to grasp the relevant aspects of a situation intuitively and directly), and can also help to formulate hypotheses regarding the nature and structure of these discriminative abilities, the way in which they interact with other cognitive capacities (the initiation and conduct of reasoning, for example), the way in which they can be refined, and so on. But if we wish to reflect on whether, to what extent, and through what means these discriminative abilities can be developed and enhanced, this type of conceptual framework proves wholly insufficient. We must supplement it with the far more precise models through which those abilities are described and studied by the cognitive sciences.

---

[7] See, on this issue, M. BRIGAGLIA, B. CELANO, *Reasons, Rules, Exception: Towards A Psychological Account*, Analisi e diritto, 2018, pp. 131–144.

The dialogue between the two conceptual frameworks – the sophisticated and rationalised common-sense psychology of philosophical and legal discourse, and the array of models from the cognitive sciences – is less straightforward than it might appear at first glance. The differences are at times so profound and yet so subtle as to create a significant risk of a breakdown in communication and misunderstandings. A process of conceptual translation and mediation is required[8]. Furthermore, a *targeted* transfer of information is required: on the one hand, the results of psychological research must be made available to philosophers and legal scholars by presenting them in a way that addresses their normative concerns; and, conversely, psychologists and neuroscientists must be asked questions on normatively relevant issues, or hypotheses suggested by legal practice or philosophical reflection, in a way that can engage with their theoretical models. This is not, of course, a new endeavour, but a path already embarked upon in many fields, including those closely related to the issue of obedience (reflection on intentional action, freedom of will, and responsibility). What needs to be done is to draw upon this existing work, directing it specifically towards the problem of responsible obedience.

One of the strengths (and most rewarding aspects) of the ABIDE project, as it has unfolded so far, has been its marked and genuine interdisciplinary nature, and the intense dialogue that has effectively taken place both between different academic disciplines (philosophy of law, philosophy of language, sociology of law, history of law, criminal law, criminal law, international criminal law, international law, comparative public law) and between different professional perspectives (academics, magistrates, prison service staff and members of the armed forces). In a (hoped-for) future development, this interdisciplinarity will need to open up to direct collaboration with the cognitive sciences.

These, however, are merely good intentions for future work. In this chapter, I shall confine myself to presenting a tiny introductory fragment, with a twofold aim: on the one hand, to give a more concrete idea of the type of conceptual translation required, and on the other, to arouse the reader's curiosity about a very recent line of research on obedience which takes its cue, albeit in a highly innovative way, from Milgram's famous experiments.

3. Milgram on the 'alienation of control'

[8] For readers interested in my work in this area, I refer you to M. BRIGAGLIA, *Foucault on Power, Law, and Society: A Reappraisal*, Routledge, London-New York 2026 (forthcoming), Chapter 7 (focusing on processes of rule-following and obedience) and *Rules: An Essay on Psychodeontics*, L'Ircocervo, 21 (2022), no. 2, pp. 210–232.

Contemporary reflection on the psychological mechanisms at work in the dynamics of obedience centres on the series of experiments conducted by Stanley Milgram in the early 1960s[9] – among the psychological experiments that have had the greatest impact on the general public, and which have sparked the most discussion and criticism in the scientific literature.

The subjects of the experiment were recruited through invitations to participate in a study on memory, offering a modest payment, advertised in newspapers, or even through invitations sent by post to previously identified individuals. Each experiment involved a 'Learner', a 'Teacher', and the experimenter. The Learner had to memorise a series of words and answer questions relating to them. The Teacher's task was to ensure the task was performed correctly and to administer an electric shock in the event of an error, with the intensity of the shocks increasing with each subsequent error. The Teacher had a control panel at his disposal, with a lever for each shock (30 shocks, ranging from 15 to 450 volts, with indications regarding the strength of the shock, from 'mild' to 'dangerous'), connected to an adjacent room where the Learner's station was located, a sort of electric chair. During the experiment, communication between the two rooms was solely verbal and not visual. An ointment was applied to the Learner, with the explanation that it served to protect against burns, and it was specified that, although the shocks might be painful, they would not cause permanent damage.

The roles of Learner and Teacher were assigned by drawing lots between two subjects, but the Learner was in fact an actor, and the draw was rigged. The real experiment aimed to test to what extent the Teacher would be willing to administer the shocks to the Learner. During the experiment, the Learner made several mistakes, which required the level of the shocks to be increased; he responded with expressions of pain and protests, which grew increasingly intense as the voltage of the administered shock rose. At 150 volts, the Learner asked for the experiment to be stopped. At 180, he said he could not take it any longer. At 270 volts, he let out agonising screams. At 300 volts, he shouted that he would no longer answer the questions, and at 330 volts the protests ceased, giving the impression that he might have lost consciousness.

If participants refused to proceed, the experimenter had four phrases at their disposal to be used in the following order: 'Please, continue', 'The experiment requires that you continue', 'It is absolutely essential that you continue', 'You have no other choice, you must continue'. The phrases were spoken in a firm tone, but without aggression. Only if the participant resisted all four prompts

---

[9] V. S. Milgram, *Behavioural study of obedience*, Journal of Abnormal and Social Psychology, vol. 67, 1963, pp. 371–378 (the article in which the results of the first series of experiments were published) and S. Milgram, *Obedience to Authority: An Experimental View*, Harper & Row, New York 1974 (the book offering a comprehensive overview of the various experiments conducted, together with an interpretation of them).

would the experiment be halted. At the end of the trial, an interview was conducted in which the experiment was explained, and it was shown that the student had suffered no harm.

Milgram classified the behaviour of those who carried on to the end as 'obedient'. In the variant of the experiment described here, the obedience rate was 62.5%, a figure much higher than expected. Another interesting aspect that emerged from the experiment is that most of the subjects who did obey right up to the strongest shocks showed, whilst inflicting pain, clear signs of distress and discomfort, both verbal and physical. To obey the order, it seemed, they had to overcome strong resistance. Obedience appeared to require confronting and resolving an inner conflict, experienced as a tragic conflict, and confronting and resolving this conflict seemed to require considerable effort.

Milgram proposed an intriguing explanation for the experiment's results. His hypothesis is that human beings have an innate tendency, shaped by natural selection as a mechanism of social coordination, to enter a peculiar cognitive and decision-making state, which he suggested calling the 'agentic state'. The agentic state is a complex condition, the main feature of which is the disposition, within a social relationship, to regulate one's behaviour according to directives from another individual of higher status, with a consequent weakening of the sense of responsibility for one's own actions. Upon entering a social context characterised by hierarchical relationships, there would tend to occur a spontaneous shift from an otherwise normal 'autonomous' state (the disposition to act on one's own initiative and on the basis of a 'free' assessment – unbound by others' prescriptions – of one's own interests) to the agentic, heteronomous state, with the inherent disposition towards obedience and a related set of attitudes – including, for example, 'attunement' to authority and the preferential attention paid to its signals. As mentioned, the tendency to enter the agentic, heteronomous state would, according to Milgram, have an innate basis – whilst also being modulated by socio-cultural factors, including ideological adherence to the authoritarian system, or beliefs regarding the scope and duties of authority.

The agentic state, as Milgram seems to conceive it, has very distinctive psychological characteristics, which make its classification ambiguous within the distinction between what is and what is not under the agent's intentional control. It is not a state of complete passivity, in which the action is not intentional but endured, as in the case of physical coercion. Nor is it a state of complete automaticity, in which the action takes place in the complete absence of intentional control, as if the body were acting of its own accord. On the contrary, in the agentic state, subjects actively control their own actions, exerting considerable effort to force themselves to obey commands whose execution elicits strong aversive reactions in them. The agentic state is not even a condition in which the action is indeed intentional, but the product of coercion or threat (*coactus tamen voluit*): one of its central characteristics is precisely that obedience to the command is not induced through coercion,

but through different mechanisms (the experimental context, as constructed by Milgram, was designed precisely to place the subjects in a position where they could choose, in the absence of coercive pressure and free from threats, whether to continue the experiment by continuing to obey, or instead to abandon it by refusing to obey). The agentic state is not merely blind, mechanical execution of the command, a complete renunciation of the critical exercise of one's own deliberative faculties, nor is it wholehearted, full adherence to the model of conduct proposed by the authority or to the objectives it pursues, as if the authority operated by virtue of attraction and emulation. On the contrary, the subjects express doubts regarding the appropriateness of the command, protest and voice dissent. Yet, when the command is (implicitly) reiterated, they obey, ending up doing something which, it seemed, they would gladly have avoided doing, and which they appeared to doubt was correct to do. Obedience, therefore, is, it would seem, 'willed', but the will itself appears to some extent to be removed from its own autonomous determination, and instead subjected to a heteronomous command. One might perhaps say that the subject retains *voluntary control* over the action, which does not occur in a purely mechanical or impulsive manner – obedience requires the subject's voluntary cooperation; it requires that the subject's capacity for voluntary control be placed at the service of the order; but, at the same time, the subject sees a higher level of voluntary control compromised to some extent: their ability to direct their own will, to choose which goal to pursue. The agentic state entails, in this sense, a bizarre 'alienation' of the subject from their own will; a curious mixture of activity and passivity – at the same time, one acts voluntarily on one's own behalf and is acted upon by another. Furthermore, the subject seems to lose an even higher level of control, their capacity for deliberative control, the ability to decide what to do and which goals to pursue, based on their own judgement of the applicable reasons. It is true that, in Milgram's interpretation, the subject recognises authority as legitimate and therefore, this seems to imply, judges that they must obey its commands – and thus, in obeying, acts in accordance with their own deliberation, obeying because they believe they have reasons to do so. But it is also true that this deliberation seems to be somewhat incomplete, the subject does not seem to adhere, with full conviction – 'in the full exercise of their rational faculties' – to the conclusion that they must obey, but seems, in some way, 'trapped' in this conclusion – they seem to act under the impetus of a cognitive and emotional pressure *that resembles a threat*, rather than acting in the light of a fully rational and convinced acceptance of a normative principle. At least, this is how Milgram seems to interpret his subjects: as individuals alienated – to some extent and at some level, difficult to specify – from control over their own decisions and actions. This state of alienation is reflected in the subjects' own interpretation of their behaviour. They seem to treat it as an intentional action, which, however, is not the result of a decision entirely of their own making, and for which they therefore do not feel fully responsible: typically,

they justify themselves by saying that they 'had to' administer the electric shock, and that 'if it had been up to them, they would not have done so'.

Milgram's concept of the agentic, heteronomous state – with its form of 'alienated' obedience, which blends activity and passivity, intentional control and heteronomy in a rather peculiar way – has been challenged by several authors, both as an explanation of the experiment's results and in general[10].

Some have argued that the obedience was not free at all, but the product of coercion[11] – or that, at the very least, it cannot be ruled out that it was (for example, because the violence to which the Learner was subjected would have created in the Teacher an unconscious expectation of suffering violence in the event of disobedience)[12].

Others have argued that this is not so much a heteronomous state as a fully convinced adherence to the experimenter's values. This criticism has been directed primarily at the extension of Milgram's theory to the crimes of the Holocaust, following Arendt's model of the 'banality of evil' – crimes committed through forms of mechanical, passive, 'alienated' obedience.

In a series of articles, for example, Alex Haslam, Stephen Reicher and colleagues have argued that the dynamics of supposed alienation of the will would instead be better explained as creative and flexible adherence to the values championed by a leader[13] . However inspired by horrific values, obedience, in this sense, would not imply any alienation of the will, but would instead be a mode of conduct that adheres entirely to the leader's will, making it its own. Whilst Milgram's obedience entails a strange loss, reduction or alienation of the self, Haslam and Reicher's obedience instead entails a reconfiguration of the self, which conforms to the leader and emulates them. (It is worth noting that the two models are not incompatible: they identify two different forms that obedience can

---

[10] An excellent review of the most recent critiques can be found in D. KAPOSI, *The Second Wave of Critical Engagement with Stanley Milgram's "Obedience to Authority" Experiments: What Did We Learn?*, Social and Personality Psychology Compass, 2022, vol. 16, no. 6, e12667.

[11] See G. PERRY, *Behind the Shock Machine: The Untold Story of the Notorious Milgram Psychology Experiments*, The New Press, New York 2013.

[12] V. D. KAPOSI, *The Resistance Experiments: Morality, Authority and Obedience in Stanley Milgram's Account*, Journal for the Theory of Social Behaviour, 2017, vol. 47, no. 4, pp. 382–401.

[13] V. S.A. HASLAM, S.D. REICHER, *Contesting the "Nature" of Conformity: What Milgram and Zimbardo's Studies Really Show*, PLOS Biology, 2012, vol. 10, no. 11, e1001426; S.A. HASLAM, S.D. REICHER, M.E. BIRNEY, *Nothing by Mere Authority: Evidence That in an Experimental Analogue of the Milgram Paradigm Participants Are Motivated not by Orders but by Appeals to Science*, Journal of Social Issues, 2014, vol. 70, no. 3, pp. 473–488; S.D. REICHER, S.A. HASLAM, A. MILLER, *What Makes a Person a Perpetrator? The Intellectual, Moral, and Methodological Arguments for Revisiting Milgram's Research on the Influence of Authority*, Journal of Social Issues, 2014, vol. 70, no. 3, pp. 393–408.

supposedly take. The point of disagreement, rather, is whether certain specific dynamics of criminal obedience are better explained by one model or the other.)

There are, however, other lines of inquiry that have instead moved in the direction of Milgram, reaching conclusions that confirm, at least in part, his hypotheses.

4. The experimental paradigm of Caspar and colleagues: sense of agency and intentional binding

In this and the following sections, I will examine some studies conducted by Emilie Caspar, Patrick Haggard and colleagues[14] based on an experimental paradigm centred on measuring the 'sense of agency' through the so-called 'intentional binding' effect.

In the first of the studies under consideration[15], Caspar and colleagues outline their general objective, as well as the conceptual framework and assumptions underpinning the experimental paradigm adopted. The general objective is to investigate the subjective experience of agents who find themselves in a condition of obedience[16] analogous to that of the Learner in Milgram's experiment: the infliction of harm (and in particular of a painful sensation) upon another innocent subject in the course of carrying out an order. More precisely, their aim is to investigate the subjects'

[14] E.A. CASPAR, F. CHRISTENSEN, A. CLEEREMANS, P. HAGGARD, *Coercion Changes the Sense of Agency in the Human Brain*, Current Biology, 2016, vol. 25, no. 5, pp. 585–592; E.A. CASPAR, A. CLEEREMANS, P. HAGGARD, *Only Giving Orders? An Experimental Study of the Sense of Agency When Giving or Receiving Commands*, PLoS One, 26 September 2018; E.A. CASPAR, S. LO BUE, P.E. MAGALHÃES DE SALDANHA DA GAMA, A. CLEEREMANS, P. HAGGARD, *The Effect of Military Training on the Sense of Agency and Outcome Processing,* Nature Communications, 2020, 11, 4366. A less technical account of her research on obedience is the very recent E. CASPAR, *Just Following Orders: Atrocities and the Brain Science of Obedience*, Cambridge University Press, Cambridge 2024.

[15] E.A. CASPAR et al., *Coercion Changes the Sense of Agency*, op. cit.

[16] Caspar and colleagues refer to the condition of obedience as 'coercion'. This choice clashes with the customary use of the terms, which link coercion to the threat of harm, and not to a mere request – as mentioned above, the central notion of obedience, the one on which both the philosophical debate and the psychological debate triggered by Milgram's experiments focus, concerns obedience rendered *in the absence of coercion.* In the last of the studies under consideration (E.A. CASPAR et al., *The Effect of Military Training*, cit.), the researchers recognised the risk of misunderstanding and explicitly stated that their use of 'coercion' differs from the customary one. To avoid misunderstandings, however, I shall describe their experiments using the term 'obedience' rather than 'coercion'.

sense of agency, defined as "the subjective experience of controlling one's actions and, through them, external events"[17].

The most obvious way to measure subjects' sense of agency is to rely on the explicit reports that the subjects themselves provide regarding whether, and to what extent, they felt they had controlled their own action and, through it, the production of a certain effect. However, the authors point out that this type of explicit measure is unreliable, given the risk that subjects' retrospective reports may be biased. For example, one might expect that those who have carried out an order to perform an action that inflicts unjust harm on someone else would tend to report a lower sense of agency due to an unconscious attempt to reduce others' disapproval.

To overcome this difficulty, Caspar and colleagues turned to an 'implicit' measure of the sense of agency, namely the perception that agents have of the time interval between an action and its effect. The use of this measure is based on a well-established psychological effect known as 'intentional binding'[18] . When an action is followed shortly afterwards by an intended effect or by an event interpreted as an effect (for example, moving a lever with a hand movement is followed 250 milliseconds later by the emission of a sound), in the agent's perception the moment of the action and that of the effect are brought closer together compared to the baseline conditions when the action is performed voluntarily, whereas they are separated when the action occurs passively and involuntarily (such as, for example, when the agent's movement is caused by transcranial magnetic stimulation of their motor cortex, or when the agent's hand is moved by the experimenter)[19] . It is in this sense that we speak of 'intentional binding': intentional action and the resulting effect are integrated by aligning them in perceived time. Following an approach already outlined in previous studies by Haggard and

---

[17] For an introduction to the concept, see P. HAGGARD, V. CHAMBON, *The Sense of Agency*, Current Biology, 2012, vol. 22, no. 10, R390; J.V. MOORE, *What Is the Sense of Agency and Why Does It Matter*, Frontiers in Psychology, vol. 7, 2012, Article 1272.

[18] P. HAGGARD, S. CLARK, J. KALOGERAS, *Voluntary Action and Conscious Awareness*, Nature Neuroscience, vol. 5, no. 4, April 2022, pp. 382–385. For an explanation of the method used to identify the effect, see M. JENSEN, S. DI COSTA, P. HAGGARD, *Intentional Binding: A Measure of Agency*, in M. OVERGAARD (ed.), *Behavioural Methods in Consciousness Research*, Oxford University Press, Oxford 2015, chap. 9.

[19] The reference condition was provided by the perceived time of the voluntary action, the involuntary action, and the sound considered in isolation. The association of the voluntary action with the effect was found to depend both on the delay in the perceived timing of the action and on the advance in the perceived timing of the effect, whilst the dissociation of the involuntary action from the effect was found to depend both on the advance in the perceived timing of the action and on the delay in the perceived timing of the effect (ibid.).

others[20] , and accepted by many[21] , Caspar and colleagues consider the production of the intentional binding effect to be a reliable measure of the sense of agency – an 'implicit' measure because it is not based on the agent's explicit evaluation, but rather on an effect of which the agent is unaware, yet which is believed to be robustly correlated with the subjective experience of having controlled the action and its effect[22].

5. Experiments and results

In the first of the studies under consideration[23], Caspar and colleagues report on two experiments designed in this way. Participants took turns playing the role of 'agent' or 'victim'. In the first group, the agent was free to choose at each turn, via a keyboard, whether or not to take money from the victim, thereby increasing their own earnings. In a second group, the agent was free to choose whether or not to administer a small electric shock to the victim, again thereby increasing their own earnings (receiving a sum of money for each shock administered). These two conditions of free choice were then compared with conditions of obedience, in which it was instead the experimenter who ordered the agent whether or not to take money from the victim or whether or not to administer the shock. Finally, a 'passivity' condition was also included, in which the experimenter physically forced the agent's hand to press the keypad, this time with a passive movement, not voluntarily controlled.

In the first experiment, the perceived time interval between action and outcome was measured. Under the obedience conditions, regardless of whether pain was inflicted or not, the interval between action and effect perceived by the agent was found to be (i) longer than that perceived under the free-choice conditions, and (ii) similar to that perceived under the passivity conditions, when the agent's hand was moved by the experimenter. The perceived time interval, as will recall, is an implicit measure of the sense of agency. The experiment, therefore, seems to suggest that, under conditions of obedience, the action is experienced in a manner akin to that in which a passive, non-voluntarily controlled action is experienced: as if one were being acted upon, rather than acting oneself.

---

[20] See, e.g., P. HAGGARD et al., *Voluntary Action*, op. cit.; P. HAGGARD, M. TSAKIRIS, *The Experience of Agency: Feelings, Judgments, and Responsibility*, Current Directions in Psychological Science, vol. 18, no. 4, pp. 242–246.

[21] See J.V. MOORE, S.S. OBHI, *Intentional Binding and the Sense of Agency: A Review*, Consciousness and Cognition, vol. 21, no. 1, 2012, pp. 546–561.

[22] For a more detailed discussion of the distinction between implicit and explicit measures of the sense of agency, see J.V. MOORE, *What Is the Sense of Agency*, op. cit.

[23] E.A. CASPAR, *Coercion Changes the Sense of Agency*, op. cit.

In the second experiment, an electrophysiological response – the so-called event-related potential (*ERP*) – stimulated by the effect of the action was measured using an electroencephalogram. Previous studies had shown a greater ERP for freely chosen actions than for the execution of instructions. The results of the second experiment by Caspar and colleagues confirmed this effect and were consistent with the findings of the first experiment: actions performed under conditions of free choice elicited a greater potential than those performed under conditions of obedience, whilst the latter, in this respect, were closer to the condition of passivity.

After the second experiment, participants were also asked to rate, as a percentage, their sense of responsibility for the action they had performed. On average, the sense of responsibility was 86.8% in the free-choice condition; 34.8% in the obedience condition; and 17.9% in the passivity condition. The participants, therefore, reported feeling much less responsible in the obedience condition than in the free choice condition, whilst the sense of responsibility felt in the obedience condition was closer to that of the passivity condition than to that of the free choice condition.

Caspar and colleagues conclude that the obedience condition has a negative effect on the sense of agency, on the subjective experience of having exercised control over one's own action and its effects: those who obey feel less in control, as if they were being acted upon passively by someone else. Justifications such as 'I was just following orders', suggest Caspar and colleagues, may therefore be more than just a way of retrospectively justifying actions that appear reprehensible. They may reflect real differences in the way the action is subjectively experienced, differences that also arise in the emotionally salient case where the agent knows they are inflicting pain on the victim, and knows this through direct experience (the participants, in fact, took turns playing the roles of agent and victim, experiencing first-hand the pain of the administered shock).

In the second of the studies under consideration[24] , Caspar and colleagues extended their investigation to the sense of agency experienced not only by those receiving orders, but also by those giving them. In the experiments described in the study, participants took turns assuming the role of Commander, Agent or Victim. In the free-choice condition, the Agent was free to decide whether or not to administer a painful electric shock to the Victim, receiving a small reward for each shock. In the obedience condition, the Commander ordered the Agent whether or not to administer the shock. At the end of the experiment, the Commander and the Agent were also asked to rate their own responsibility on a percentage scale. For our purposes, it is not necessary to dwell on further details of the experiments and their differences; it suffices to describe the main results obtained. Regarding the comparison between the free-choice condition and the obedience condition, the results of the

[24] E.A. CASPAR et al., *Only Giving Orders*, op. cit.

previous study were confirmed – with the addition that in the new study the effect occurred in the absence of an institutional role held by the person issuing the orders (each of the study participants in turn, rather than the experimenter, as in the first study), suggesting that the distinctive features of the obedience situation are activated, to a significant extent, even outside a recognised institutional context. With regard to the command condition, however, new and interesting data emerged. As expected, participants considered themselves more responsible in the free-choice condition (89%) and in the command condition (81%) than in the obedience condition (41%). Less predictably, however, even in the command condition, a greater interval was observed between the perceived time of the action and its effect than in the free-choice condition, and closer to that recorded in the obedience condition than in the free-choice condition. The lengthening of perceived time, it will be recalled, is considered an implicit indicator of a *reduction* in the sense of agency. What these results therefore seem to suggest is that those who command someone else to perform an action feel less in control of the production of the final effect than those who produce the effect directly through their own action, and that this reduction in the sense of control approximates that characteristic of the obedience condition. To explain why those in the commanding position, despite the reduction in their sense of agency, still considered themselves responsible for the effects of the commanded action, Caspar and colleagues hypothesised that the assessment of responsibility was modulated not only by the sense of agency experienced, but also by an internalised normative model according to which one *ought* to feel responsible for the command.

In the third of the studies under review[25], Caspar and colleagues set out to investigate the influence of military training on the sense of agency in the obedience condition. The experimental design was similar to that of the experiments in the first study. In the free-choice condition, the agent was free to choose whether to administer electric shocks to the victim, receiving a reward for each shock administered. In the obedience condition, however, they were ordered to do so.

In an initial experiment, the responses of civilians and junior cadets were compared. In the obedience condition, the junior cadets showed an increase in the perceived interval between action and effect, indicative of a reduction in the sense of agency, similar to that of civilians. Unexpectedly, however, they also showed, compared to civilians, an increase in the perceived interval – indicative of a reduction in the sense of agency – even in the free-choice condition. The effect of partial alienation from control over one's own actions, characteristic of the obedience condition, seemed to extend even to freely made choices. Another interesting finding concerns the irrelevance of the institutional role of the person issuing the orders. The subjects had in fact been divided into two

[25] E.A. CASPAR et al., *The Effect of Military Training*, op. cit.

groups. Whilst in one group the orders were given by a military officer with the rank of captain, in uniform, in the other group they were given by a researcher, in civilian clothes. Despite this difference, the results of the two groups showed no significant differences, confirming that the peculiarities of the obedience condition are at least partly independent of institutional contexts.

In a second experiment, the responses of privates (corresponding to troop soldiers) with approximately five years' service and senior cadets with an average of five years' training and the rank of lieutenant were investigated. The data relating to senior cadets confirmed those of the previous experiment, showing here too an increase in the perceived interval between action and effect, indicative of a reduction in the sense of agency, both in the obedience condition and, compared to the data relating to civilians, in the free choice condition. Similar data were recorded for privates. In senior cadets, however, whilst the reduction in the sense of agency in the obedience condition was similar to that of the other categories (civilians, junior cadets, privates), the reduction in the sense of agency in the free choice condition disappeared, returning to the level of civilians.

The data relating to senior cadets and privates, the researchers suggest, raise the question of whether prolonged training in carrying out orders – typical of a position of subordination within a hierarchical structure – may have a negative impact on the perception of one's own actions, reducing the sense of being in control of the action and its effects not only when obeying, but also when choosing freely. Constantly receiving orders can create a new normality in which the experience of acting voluntarily comes to resemble the partially alienated experience of following orders. Data on senior cadets, on the other hand, seem to indicate that this effect can be countered by training in the assumption of role-related responsibilities, allowing for a note of optimism regarding the possibility of developing a culture of responsibility within hierarchical organisations.

6. Conclusions and some comments

In Caspar and colleagues' interpretation of their series of experiments, there are essentially two key conclusions that can be drawn. (1) The first, explicitly stated, is that in the obedience condition (and in the command condition) there is a very significant reduction in the sense of agency, which means that the experience of obedient action comes close to that of a passive, involuntary action, controlled by external forces. (2) The second, less explicit but clearly implied by their considerations, is that the reduction in the sense of agency described in (1) is accompanied by some reduction in the agent's capacity to intentionally control their own action, resulting in a diminished ability to resist orders, including orders that the agents themselves would deem illegitimate, criminal or immoral – it

is only under this condition that the experiments can have the value the authors attribute to them with regard to the problems and dangers of obedience. Both of these conclusions require some comment.

Let us begin with the first conclusion, concerning the reduction in the sense of agency. As we have seen, Caspar and colleagues distinguish between an implicit measure – the effect of *intentional binding* – and an explicit measure – subjective reports – of the sense of agency. It is not entirely clear, however, whether they intend these to be two measures of the *same mental event*, or rather measures of *two distinct*, albeit related, *mental events*. The second interpretation seems preferable to me. The sense of agency would therefore have at least two dimensions, or layers, or levels: the first dimension (let us call it SdA-1) is measured by the intentional binding effect, whilst the second dimension (let us call it SdA-2) is measured by subjective reports. These two dimensions of the sense of agency appear, first and foremost, to have distinct temporal locations. SdA-1, measured by the effect of intentional binding, is experienced *contextually*, in the immediate temporal proximity of the action and the perception of its effects. SdA-2, on the other hand, consists of the subject's recall or reconstruction of the experience of agency at a later time, further removed from the temporal context of the action. This recall may be modulated not only by the memory of SdA-1, but also by other factors, including *bias*. It should be emphasised that the studies by Caspar and colleagues primarily concern SdA-1. SdA-2 is not investigated. Rather, a dimension likely connected to it is investigated, namely the *responsibility* that subjects explicitly attribute to themselves for producing the effects of their actions. In addition to temporal placement, there may be another important difference between SdA-1 and SdA-2. SdA-2 seems to consist of, or incorporate, a genuine *judgement*: a logically structured thought that explicitly attributes a subjective experience to oneself. SdA-1, on the other hand, might be, rather than a logically structured thought , a sort of elementary feeling, or even a preconscious event. The idea that the sense of agency has at least two components, a 'feeling of agency' and a 'judgement of agency', is, in fact, a prominent and influential idea in the specialist literature[26]; whilst the idea that the intentional binding effect measures precisely the feeling of agency has been accepted in an article co-authored by Haggard[27]. It is therefore not unfounded to interpret the studies by Caspar, Haggard, and colleagues in this way. Whatever its nature, however, Caspar and colleagues believe that SdA-1 is (i) a real mental event, objectively measurable through the

---

[26] V. M. SYNOFZIK, G. VOSGERAU, A. NEWEN, *Beyond the Comparator Model: A Multifactorial Two-Step Account of Agency*, Consciousness and Cognition, vol. 17, 2008, pp. 219–239. The 'feeling of agency' is defined as "the non-conceptual, low-level feeling of being the agent of an action", whilst the 'judgment of agency' is defined as "the conceptual, interpretive judgment of being an agent."

[27] J.W. MOORE, D. MIDDLETON, P. HAGGARD, P.C. FLETCHER, *Exploring Implicit and Explicit Aspects of Sense of Agency*, Consciousness and Cognition, 21, 2012, pp. 1748–1753.

intentional binding effect; (ii) for which neural correlates can be identified[28] ; and (iii) that it has a specific functional role: at some level, SdA-1 serves the subject – or, if one prefers, some mechanism in their brain – to recognise the action and its effect as subject to their own voluntary control – it is a marker that conveys internal information regarding the voluntary control of the action and its effect.

It is presumably on this very point that the authors' second conclusion rests, which echoes Milgram's hypothesis: when the effect of intentional binding is reduced, this 'alienates' the subject from their own will, preventing them, at a fundamental level, from recognising the action and its effect as being subject to their own voluntary control. This, in turn, makes it somewhat more difficult for the subject to exercise higher-level control over the action, *thereby facilitating* the execution of the order. The condition of obedience, therefore, fosters a state of alienation from control over the action, which facilitates the execution of the order.

The authors do not specify what form of higher-level control is hindered by obedience. But one might speculate that, if their conclusions are on the right track, it is precisely the type of control I have called 'deliberative control': the monitoring of the action on the basis, amongst other things, of its conformity to high-level normative schemes (accepted norms and values, pursued goals, and the like). The agent, not fully recognising themselves as being in voluntary control of the action and its effects, would be hindered in exercising deliberative control over it – it would be more difficult for them to monitor the correctness of the action on the basis of normative standards, and/or to intervene, in the event of non-compliance, to prevent the action. In short: the mechanism of reducing *intentional binding* would help to hinder and reduce the exercise of deliberative control over the content of the order, thereby making its execution more 'automatic' in this sense, and increasing its likelihood.

Caspar and colleagues also venture a comment on the legal regulation of obedience: "The defense of 'only obeying orders' is often treated with suspicion in law because of the clear secundary gain associated with denying responsibility. However, our result suggests that primary feelings and neurophysiological processing of agency are indeed reduced by coercion [i.e., in the obedience condition]"[29]. Quite rightly, the authors do not go so far as to suggest that their studies call for a *radical* rethinking of criminal responsibility for actions carried out in execution of orders (recognising it, in general, as a ground for justification, or as a condition that reduces the subject's culpability or blameworthiness). At most, the studies support the conclusion that the condition of obedience *hinders*

---

[28] In a further study, the researchers demonstrated that the neural correlates of the sense of agency are also reduced in a state of obedience (E. CASPAR, F. BEYER, A. CLEEREMANS, P. HAGGARD, *The Obedient Mind and the Volitional Brain: A Neural Basis for Preserved Sense of Agency and Sense of Responsibility Under Coercion*, PLoS ONE, 16, 10, 2021, e0258884.

[29] E.A. CASPAR, *Coercion Changes the Sense of Agency*, op. cit.

the exercise of deliberative control over the order, identifying one of the mechanisms responsible for this difficulty – which in no way implies that the condition of obedience renders the exercise of deliberative control over the content of the order *impossible*, or renders it so difficult as to make the legal requirement to exercise such control and the attribution of liability in the event of failure appear unreasonable.

However, this necessary caution does not detract from the fact that, should the conclusions be confirmed, they would be highly significant for the issue of responsible obedience. The mechanism identified could nevertheless provide a useful clue, alongside other factors, for identifying those cases in which the capacity for deliberative control of the individual who committed a crime whilst carrying out an order was so compromised as to justify a reduction in their criminal liability. Above all, however, it could help us understand how *to build and facilitate* responsible obedience, by designing training methods and decision-making contexts that help restore the sense of agency in the person who obeys, thereby facilitating their exercise of deliberative control.