Stanley Milgram’s Obedience Studies: A Critical Review of the Most Influential Explanatory Accounts

Nestar Russell and Raphaël Künstler

Electronic version
URL: https://journals.openedition.org/philosophiascientiae/4328
DOI: 10.4000/11pu0
ISSN: 1775-4283

Publisher
Éditions Kimé

Printed version
Date of publication: June 2, 2024
Number of pages: 3-31
ISBN: 978-2-38072-141-6
ISSN: 1281-2463

Electronic reference

The text only may be used under licence CC BY-NC-ND 4.0. All other elements (illustrations, imported files) are "All rights reserved", unless otherwise stated.
Stanley Milgram’s Obedience Studies: A Critical Review of the Most Influential Explanatory Accounts

Nestar Russell
University of Calgary (Canada)

Raphaël Künstler
CLLE–ERRAPHIS, Université de Toulouse (France)

Abstract: In 1960 Stanley Milgram wondered if ordinary people would, as many Germans did during the Holocaust, obey higher orders to harm innocent people. Soon after, he ran the New Baseline experiment: at the behest of a scientific authority, 65 percent of subjects inflicted what appeared to be potentially lethal shocks on another person with a mild heart condition. Although, to date, there is no widely accepted explanation that can account for this finding, three theories continue to attract a disproportionate amount of attention: the Incredulity Hypothesis, Agentic State, and Engaged...
Followership. The aim of this paper is to present an overview of how Milgram invented his basic procedure and then use the insights gained from his journey of discovery as a foundation from which to critically review these three explanatory accounts. Succeeding this critical review, a lesser well-known theory will be presented; one that is supported by Milgram’s original results.

*A piece of literature makes its way into canon based largely on the meaning it imparts in our lives. Milgram’s experiments are indisputably in the canon. And yet, no one can agree on the theme [...] What message has Milgram sent us, in what sort of bottle, on which sea?*

Lauren Slater [2004, 61]

1 Introduction

In 1960, the American social psychologist Stanley Milgram, both inspired and terrified by the Holocaust, wondered if ordinary (American) people would follow a higher authority’s orders to harm an innocent person. In his *New Baseline* experiment, undertaken at Yale University, the subject was informed by an actor playing the role of a social science “experimenter” that they were to participate in an experiment purportedly exploring the effects of punishment on learning. The subject was allocated the role of “teacher” and was taken to a small room where they observed the experimenter firmly strap another actor fulfilling the role of “the learner” into an electric chair. In conversation, the learner warned all present that they had a mild heart condition. After reassuring the learner that although the shocks “may be painful they’re not dangerous,” the experimenter, with subject in tow, left the small room and entered a larger adjoining room where the subject was seated in front of a shock generator: an electrical device with 30 switches that increased in 15-volt increments ranging from 15 to 450 volts.

The experimenter then instructed the subject to ask the learner a list of word pair questions, which the learner purposefully answered incorrectly with regularity. The experimenter instructed the subject to increase the intensity of the “punishment” by 15 volts for every incorrect answer received. No shocks were really administered, but if the subject hesitated to do as they were told, the experimenter urged them to follow his shock-inflicting instructions
by deploying a set of increasingly insistent prods. If the subject complied, at the 150-volt switch, the learner exclaimed his heart was bothering him. During the second half of the experiment, the learner complained further about his heart and then started engaging in agonized screams until, across the last third of the remaining shocks, he went silent, implying something had gone terribly wrong. To this, the experimenter urged the subject to treat all subsequent unanswered questions as incorrect and continue inflicting further shocks. Milgram’s research question was: would the subject comply or defy the experimenter’s insistence that they continue to inflict what appeared to be excruciating shocks on a vulnerable person?

Essentially, this procedure forced subjects into making a choice: side with the learner and stop inflicting painful shocks, or side with the experimenter and collect further data by inflicting more shocks. The subject could not simultaneously appease both parties. They were faced with resolving a dilemma: was it more important to help the experimenter or the learner? Milgram argues this dilemma has a moral dimension: when observers of the experiment are asked “what constitutes appropriate behavior in this situation, they unfailingly see disobedience as proper” [Milgram 1974, 6]. Disobedience was proper because of the experimenter’s “[…] fundamental breach of moral conduct to hurt another person against his will” [1974, 41]. For outsiders looking in, the choice is obvious: it is more important to help an innocent person with a mild heart condition avoid intense shocks than it is to collect data. But herein lies the attraction of laypersons and scholars alike to this research: refusing to inflict shocks appears to be the morally correct course of action, yet the New Baseline experiment showed that 65 percent of subjects instead chose to inflict what they were led to believe were three consecutive 450-volt shocks on an innocent person. Thus, an intriguing question arises: why did most subjects repeatedly pursue what outsiders unanimously agree was the wrong choice?

Over the last half century or so, a range of theories purporting to explain Milgram’s counter-intuitive results have been proposed. Although some explanatory accounts have generated a greater following than others, scholars specialising in the Obedience Studies have failed to converge in support of any one of them. That said, the central aim of this paper is to present and critically review the three most prominent explanatory accounts for Milgram’s baseline results: the Incredulity Hypothesis, the Agentic State, and Engaged Followership. The final section of this article will briefly present a less known theory, the validity of which, it will be argued, is bolstered by Milgram’s own results. Before presenting these theoretical accounts, however, it is important to note that one potentially fruitful strategy to better understand the inner workings of the Obedience Studies is to delineate their start-to-finish invention. Doing so, we believe, will provide a strong base from which to critically assess the quality of all these theoretical accounts. Thus, what follows will first present an abridged overview of Milgram’s inventive journey of discovery.
2 Invention and key findings

During the Nazi war crimes trials, many ordinary Germans justified their harmful actions by arguing that they were just following higher orders. Milgram therefore wondered what would happen if he ran a social science experiment where an authority figure instructed ordinary people to follow orders to hurt another person. For this experiment to garner scholarly attention, Milgram knew it needed to obtain eye-catching results—nobody would be surprised by an experiment that generated a low rate of harmful obedience. He therefore set out with a preconceived goal: the official baseline experiment had to “maximize obedience [...]” [Milgram cited in Russell 2011, 158]. Milgram, however, did not have an experimental procedure capable of producing this result, so he had to invent one.

His first attempt at inventing a procedure capable of maximising obedience was rudimentary. Initially he planned to hire an actor to play an authority figure who, for some reason, required the subject to obey his commands to physically assault another person. As Milgram noted in one document (circa 1960):

We can then start out by giving the subject commands from the lower end. (Tap him.) And gradually proceed to more intensive commands. (Slug him.) Some kind of scale of this sort is needed. A second requirement [...] is to give the command some institutional justification. [Milgram cited in Russell 2018, 48]

Inherent in this idea is the manipulative “foot-in-the-door” technique: persons are more likely to agree to a significant request if it is preceded by a comparatively insignificant request [Freedman & Fraser 1966]. Milgram later termed his inclusion of manipulative techniques like this as Binding Factors, which are powerful bonds that can entrap a person into engaging in actions they would prefer not to do [1974, 148]. But were “intensive commands” to engage in an aggressive and intensifying “scale[d]” physical assault likely to maximise obedience? Two problems with this idea probably became apparent. First, putting aside the ethical difficulties of running an experiment where a person was physically beaten(!), surely subjects would feel too closely connected to the order’s harmful end result. Second, subjects would surely need a reason—an “institutional justification”—for instigating the assault. Subsequent procedural changes indicate Milgram likely sensed then resolved both problems. That is, Figure 1 (below), a document written by Milgram, indicates subjects were to harm another person by inflicting intensifying electrical shocks. This document also identifies the “institutional justification” Milgram intended to use to encourage harm-infliction: before starting, subjects were to accept a “pledge to obey”: 
Figure 1: A document from Milgram’s personal archive titled “Studies in Obedience...” [cf. Russell 2018, 49]

The contents of this document are as follows:

Studies in Obedience

1. Wa[ɪ]ver of responsibility—from experimenter—For Germa[n]y
2. Panel [Sketch of shock generator with the final shock labeled “LETHAL”]
3. The War Situation—2 naive S[ubject]’s. One must shock the other – 1 way switch. Can be controlled by E[xperimenter].
4. Working in teams:

5. The Pledge. Subjects pledge to obey. Because of certain possible hazards, the S. must adhere carefully to the instructions of the Experimenter.

In this document Milgram seems to have incorporated into his emerging procedure what he understood to be the Nazi regime’s most effective techniques of manipulation: displacement of responsibility, hierarchical obedience to authority, a pledge to obey harmful orders, taking small steps toward a radical outcome, all in the context of war.

But the document’s most eye-catching feature is Milgram’s idea to utilize a shock generator as the means of harm infliction. This device was an example of what he later termed a Strain Resolving Mechanism. Strain Resolving Mechanisms help reduce the stress people normally experience when inflicting harm [Milgram 1974, 153–164]. That is, a person who presses switches that harm another person will experience less stress than had they physically beaten them with their fists.

However, one weakness with his above research idea was that surely the “pledge to obey” was too closely associated with the Nazi regime whereby Milgram’s inclusion of it risked alerting subjects to the study’s actual research question. To increase the research idea’s viability, Milgram required a more veiled justification for harm infliction. Indeed, by October 1960 Milgram had discarded his pledge to obey in favour of an more opaque—better disguised—justification for harm infliction: the subject, in the role of a “teacher”, was to inflict shocks on a “learner” to help determine if punishment improved learning [Russell 2018, 56–59]. In doing so, however, Milgram added another Binding Factor to his procedure: to help advance scientific knowledge on the topic of punishment’s effect on learning, subjects would be pressured into inflicting every shock. With his shock machine and a less transparent justification for shock infliction, Milgram envisioned two essential components that moved him closer to preconceived goal achievement.

To see if his inchoate procedure worked, Milgram’s students ran the first pilot trial and to his astonishment, most subjects inflicted every shock. Despite his surprise, Milgram remained critical of his student’s “not very well controlled” pilot [Milgram cited in Blass 2004, 68]. More specifically, because of the students’ disjointed procedure, fake-looking shock generator, and weak acting skills [Russell 2018, 64], Milgram suspected some subjects completed the experiment because they did not believe the learner was being harmed. Milgram sought to overcome these weaknesses [Blass 2004, 75], [Russell 2018, 68, 76].

It also transpires that Milgram used the pilots as a means of increasing the emerging procedure’s efficacious power. He closely observed the defiant pilot subjects’ strategies to prematurely ending the experiment and, after doing so, he then tried to envision counter-escape strategies—binding or
strain-resolving techniques—that the experimenter could deploy against future subjects who might try to use similar tactics. Doing so better ensured that during the official experiments reluctant subjects were more likely to continue inflicting further shocks. One example of Milgram utilizing this technique comes from a “Report on Pilot Res[e]arch” dated 4 December 1960, where he noted some subjects informed the experimenter it was the voluntary learner’s prerogative to “leave whenever he wants to” [Russell 2018, 64]. Thus, these pilot subjects’ awareness of a person’s right to withdraw from an experiment emboldened them to resist the experimenter’s demands that they continue. To block off this potential escape route, Milgram—clearly drawing on an idea mentioned in Figure 1—proposed “the following change”: the experimenter should respond to such future resistance by stating “I [h]ave responsibility for this situation [...]. Proceed with the next question” [Russell 2018, 65]. By the official experiment this statement was changed to “I’m responsible for anything that happens to him. Continue, please” [Milgram 1974, 74]. Milgram’s change here resulted in the addition of another Strain Resolving Mechanism to the emerging procedure: if future subjects accepted the experimenter’s offer, they could, if they so chose, displace all responsibility for shock infliction on to him (thereby potentially reducing their own strain over shock infliction).

Milgram’s varied techniques of discovering, then adding an array of Strain Resolving Mechanisms and Binding Factors to his emerging procedure—many of which are not mentioned here [see Russell 2011, 2018]—in part explains how he managed to construct a powerful official baseline procedure. Basically, Milgram—in the role of a goal-directed social engineer—kept adding Binding Factors and Strain Resolving Mechanisms to the emerging procedure, which cumulatively helped him achieve his preconceived goal, that is, maximization of the official baseline’s completion rate. The high completion rate of his final pilot trial signaled to Milgram he was finally ready to run the first official baseline. He then ran his first official experiment: the Remote Feedback condition, where the learner, who could not be seen or heard, suddenly kicked the adjoining wall at the 300-volt switch and thereafter went silent. This experiment obtained a 65 percent completion rate. Soon after running this first baseline, Milgram lost access to his laboratory space and was moved to a new location (which put pressure on him to start over). Around the same time, he came to suspect that in the newly allocated laboratory he could, on starting over, probably run a new yet far more disturbing (impressive?) baseline condition, which he suspected would still obtain a high completion rate. The New Baseline (or “cardiac” condition), which, as earlier described, involved a screaming learner distressed by his heart condition. The New Baseline also obtained a 65 percent completion rate.

After Milgram had achieved his preconceived goal of obtaining a high baseline completion rate, he then pursued his strategy for theoretical development. That is, to discover why most subjects completed the New Baseline, he planned to run a score or so of slight baseline variations, against which he
could test both earlier subjects’ post-experimental justifications for completing and the accuracy of various hypotheses as they came to him. Across the many slight baseline variations, he planned to test and eliminate these justifications and hypotheses until the most accurate account emerged. For example, some subjects claimed they only completed the experiment because it was conducted at Yale University—a respected research institution. To test the validity of this explanation, Milgram ran the Institutional Context condition: a baseline replication run in a shabby Bridgeport office under the fictitious title of Research Associates of Bridgeport. This slight variation, however, obtained a 47.5 percent completion rate; a result “not significantly lower than that obtained at Yale”, thus eliminating this potential explanation for the baseline result [Milgram 1974, 69]. Milgram also wondered if inherent human aggressiveness was responsible for the baseline result. Consequently, he ran a variation termed the Subject Chooses Shock Level condition where subjects were instructed to shock at any intensity of their choosing. Most subjects—97.5 percent—chose to inflict low-intensity shocks. Because nearly all subjects, on their own volition, chose not to seriously hurt the learner, inherent human aggressiveness did not generate the high baseline completion rate.

The theory Milgram came to favour was that ordinary people seem to have a strong proclivity to obey what they believe to be legitimate commands. He favoured this account because when subjects were given orders to inflict shocks, as in the New Baseline, 65 percent completed the experiment. But when no orders were given, as in the Subject Chooses Shock Level condition, only 2.5 percent completed the experiment. The 62.5 percent difference between both experiments bolstered Milgram’s presupposition that higher orders seemed to be centrally involved in generating the baseline result.

However, some of the results from his baseline variations conflicted with this conclusion. One of several possible examples was the Teacher in Charge condition. Before the start of this variation the experimenter stated that he had to leave the laboratory, but as he was departing, he told an actor-teacher to invent a punishment regime, which the other subject-teacher was to implement. The actor-teacher decided to increase the intensity of punishment by 15 volts for each incorrect answer, and then instructed the subject to fulfill the role of shock inflictor. In this condition, 55 percent of subjects did as the man of equal status told them and inflicted every shock. Milgram was confused: “we were surprised aond [sic] somewhat troubled that 11 [out of 20 subjects] should be full[y] obedient under common man authority” [Milgram cited in Russell 2018, 99]. Milgram was troubled because this result conflicted with his favourite theory that obedience to an authority caused most subjects to complete the New Baseline condition. Thereafter, Milgram’s confidence in developing a comprehensive theory of obedience to authority declined [Milgram cited in 2018, 99]. In short, Milgram was baffled by his occasionally contradictory baseline variation results, which saw him, across most of the 1960s, abandon his goal of developing a comprehensive theory [2018, 101].
In December 1961 Milgram submitted for consideration his first Obedience Study journal article. This paper described and presented the results of the Remote-Feedback condition (the first baseline he, as mentioned, later rejected in favour of the New Baseline; whereby the latter became the procedure that all subsequent slight variations were modeled upon). However, because in this article Milgram failed to provide a concise theoretical account for the result, reviewers twice recommended its rejection [Blass 2004, 114], with one, Edward E. Jones, adding that the 65 percent completion rate was, in support of the above overview of Milgram’s inventive journey, at best a “triumph of social engineering” [Parker 2000, 112, cited in Russell 2018, 72]. With no theory, Milgram [1963, 377] then seems to have appeased his reviewers by adding 13 “features” inherent to the procedure that he thought helped explain why most subjects completed the experiment. In 1963, this article was accepted for publication. Milgram’s research journey shows he preferred—probably with concision and simplicity in mind—a mono-causal rather than a multi-causal explanatory account for his high baseline completion rate. But, as mentioned, across most of the 1960s he failed to develop one.

Interestingly, Burger [2009, 3], who undertook a partial replication of the Obedience studies, came to a similar multi-causal conclusion to Milgram [1963] when he (Burger) presented “four features of the situation that likely contributed to the high rates of obedience [...]”: “Obedience to Authority”, “Gradual Increase in Demands”, “Limited Sources of Information in a Novel Situation” and “Responsibility Not Assigned or Diffused...” Other scholars have come to similar multi-causal conclusions, see [Bauman 1989, 151–164], [Hollander & Turowetz 2017, 671–672], [Miller 2014, 560–564]. The ensuing array of terms (“factors” and “mechanisms”)—many possible explanatory accounts—reflects a statement Milgram made in his research notes: “[It is not that] an explanation is lacking; but rather [...] there are so many to choose from” [Stanley Milgram Papers, Box 46, Folder 175, Titled: “Notes: Proximity”]. Here Milgram is alluding to a weakness with the multi-causal explanatory accounts: they inject greater ambiguity (or complexity) over the question of why exactly subjects completed the baseline. That is, their many possible explanations collectively blend back into a rather unsatisfying hazy mystery that only returns one back to the original question: why did they do it?

Milgram’s above inventive journey may, in fact, have generated this failure, across the 1960s, to produce a singular mono-causal, over his above vague multi-causal, theoretical account. That is, as shown, when Milgram constructed his basic procedure, he added so many different manipulative techniques of social engineering—all those Strain Resolving Mechanisms and Binding Factors—to the baseline procedure that his limited score of experimental variations proved incapable of isolating all their independent influences. That is, there are many different potential explanations because there were many individual Strain Resolving Mechanisms and Binding Factors inherent within the procedure.
Although Milgram, Burger, and many others have presented multi-causal accounts to help explain the Obedience Studies, more ambitious mono-causal theories have been proposed. These mono-causal theories, if accurate, have a major advantage over the multi-causal accounts: they provide greater precision and parsimony over the question of why subjects completed the New Baseline. This potential for greater precision and parsimony essentially renders mono-causal theories the holy grail of explanatory accounts, and thus, compared to the unsatisfying multi-causal theories, they tend to attract enormous scholarly attention. Indeed, this is, at least in part, why the three most prominent theoretical accounts of the Obedience Studies are all mono-causal. The potential of these three most prominent theoretical accounts to shed light on why exactly subjects completed the New Baseline renders it important that their validity be assessed. Doris, Niemi et al. [2024] term the first of these three mono-causal accounts the Incredulity Hypothesis.

3 The Incredulity Hypothesis

Milgram [1972, 139] argued that successfully deceiving most subjects into believing that the shock generator was seriously harming the learner was absolutely “critical” for the methodological strength (internal validity) of the experiments. Deception was important because if most subjects did not believe the learner was being harmed, then there would not have been any “conflict” of conscience [Milgram 1963, 378] over whether or not to continue helping the experimenter collect his data. It is therefore widely accepted that “the entire [methodological] foundation of the obedience research rests on the believability of the victim’s increasingly mounting suffering” [Miller 1986, 143]. Successfully deceiving most subjects was also of great importance to Milgram because it made the next step in his research possible: generalizing his findings to the outside world (ecological validity); most controversially, to the Holocaust [Milgram 1963, 371].

But if critics could show the basic procedure failed to successfully deceive most subjects, then Milgram could no longer, as he frequently did, generalize from his findings to the outside world. Put differently, one cannot generalize to, say, the Holocaust or elsewhere, from a methodologically weak foundation. And over the past half century or so numerous scholars have challenged Milgram’s assertion that he successfully deceived his subjects into believing they were inflicting dangerous shocks.

The first scholars to challenge Milgram on these grounds was Orne & Holland [1968] They argued that the baseline result can be explained by the obedient subjects’ accurate suspicion that the experiment was a ruse. Subjects believed it to be a ruse because they could not imagine that any researcher or prestigious institution would allow an innocent person to be exposed to danger. Therefore, subjects concluded that, despite evidence to the contrary,
no serious harm would come from their completing the experiment: doing so would be “all right” [Orne & Holland 1968, 287].

Yet if subjects did not believe the learner was being shocked, why did most complete the experiment? Orne & Holland argued that most subjects completed the baseline experiment because they were influenced by demand characteristics. Demand characteristics are when subjects attempt to detect the meaning and purpose of an experiment and, on doing this, they then engage in behaviors they think are likely to please the researcher by confirming their probable hypothesis. Thus, for Orne & Holland, most of Milgram’s subjects accurately detected the experiment’s covert aim: would they follow malevolent orders to harm an innocent person. Subjects, having established that the learner was not really being harmed, then proceeded to please Milgram by inflicting every shock asked of them whilst feigning explicit indicators of nervousness. Therefore, those subjects who completed the experiment did so to help Milgram achieve his goal of obtaining a high completion rate whilst also knowing that no harm would come from their doing so.

So how valid is Orne & Holland’s argument? In conflict with their accusation that demand characteristics generated the baseline result, Milgram’s inventive journey illustrates that when the Nazi-sounding “pledge to obey” justification for harm infliction was substituted in favour of an experiment that was (apparently) exploring the effects of punishment on learning, he clearly tried to hide from subjects his study’s actual objective of determining if ordinary people would follow a malevolent authority’s instructions. This inventive journey also shows that Milgram actively went to great lengths to protect his study from demand characteristics (recalling his skepticism with the first student-run pilot and, thereafter, the numerous ways he attempted to strengthen the basic procedure). What convinced Milgram he had successfully deceived most subjects was, of course, their signs of “extreme tension”: sweating, trembling, stuttering, groaning, biting their lips and digging their nails into their flesh [Milgram 1963, 375]. But one particularly convincing line of evidence detracting from Orne & Holland’s thesis was that if the obedient subjects knew the shocks were fake, why after completing the New Baseline experiment did 73 percent of them decline an opportunity to experience a 450-volt shock [Milgram 1974, 57]? Milgram showed no patience for Orne & Holland’s thesis, arguing their,

[...] suggestion that the subjects only feigned sweating, trembling, and stuttering to please the experimenter is pathetically detached from reality, equivalent to the statement that hemophiliacs bleed to keep their physicians busy. [Milgram 1972, 140]

Eckman [1977, 94] agreed: “when one reads the actual transcripts of Milgram’s subjects’ verbal behavior, it is hard to conclude it was all a put-on. There was just too much conflict and stress.” Thus, “[t]o invoke the charge ‘demand characteristics’ against Milgram’s work is foolish” [1977, 95].
Although there are strong grounds for dismissing the validity of Orne & Holland’s central argument that demand characteristics generated Milgram’s results, their issue of trust—that obedient subjects did not believe the shocks were dangerous and therefore that Milgram failed to deceive them—has had a powerful influence on the subsequent literature, see [Gibson 2019, 38–39], [Perry 2013, 173, 258], [Perry, Brannigan et al. 2020]. Much of this influence, however, traces not to Orne & Holland, but to Don Mixon. Across a series of publications in the 1970s and 1980s, Mixon provided a new and more persuasive explanation for the subjects’ stressed reactions, which, as a result, elevated Orne & Holland’s issue of trust to new and more influential heights.

Mixon set out by arguing that subject displays of stress and anxiety were, contra Orne & Holland, genuine. However, this stress was not due to, as Milgram argued, a conflict of conscience, but because of their exposure to an ambiguous situation where the information coming from the experimenter and learner was contradictory: the shocks were apparently harmful, but not dangerous; the experimenter was calm, but the learner was screaming in agony. Milgram’s ambiguous procedure therefore generated confusion among his subjects. As Mixon observed, “No wonder many subjects showed such stress. What to believe? The right thing to do depends on which actor is believed” [Mixon 1989, 33]. Then, according to Mixon, because “increasingly large chunks of the social and physical world that we live in can be understood only by experts” [1989, 35], most of the obedient subjects chose to resolve the stressfully ambiguous situation confronting them by trusting the authority figure’s word: he said that although the shocks “may be painful” they would “cause no permanent tissue damage [...]” [Milgram 1974, 56]. Thus, for the obedient subjects, the study must have been benign.

In support of Mixon’s claims, the above invention of the Obedience Studies illustrates that during the pilot studies Milgram indeed purposefully injected ambiguity into his emerging basic procedure. For example, Milgram changed the designated title of the last button on his proposed shock machine from “LETHAL” (Figure 1) to “XXX” (official shock generator [Milgram 1974, 28]), presumably because he sensed the definitiveness surrounding a word like “LETHAL” might detract from his preconceived goal to “maximize obedience [...]”. And there are other examples were Milgram removed clarity and replaced it with ambiguity [see Russell 2018, 123]. Thus, there is support for the claim that Milgram purposefully created an ambiguous and thus confusing situation where, in pursuit of stress reduction, subjects might have decided to side with the only expert present. Again, many contemporary scholars, like Burger, see merit in Mixon’s argument:

When you’re in that situation, wondering, should I continue or should I not, there are reasons to do both. What you do have is an expert in the room who knows all about this study, and presumably has been through this many times before with many participants, and he’s telling you, [t]here’s nothing wrong. The
reasonable, rational thing to do is to listen to the guy who’s the expert when you’re not sure what to do. [Burger cited in Perry 2013, 359]

Although Mixon’s argument is both different to and more convincing than Orne & Holland’s, the conclusions of both groups of scholars on the issue over whether or not Milgram successfully deceived his obedient subjects are the same. To clarify, because, for Mixon, obedient subjects believed the “expert” experimenter when he said the shocks were not dangerous, what is implied is that Milgram’s all-important attempts to deceive them into believing they were seriously harming the learner must have failed. Therefore, both Orne & Holland and Mixon agree: obedient subjects trusted that the experimenter would not allow the learner to experience dangerous shocks and this belief—that Milgram’s attempts at deception failed—mono-causally explains why they decided to complete the baseline experiment.

More recently, Perry, Brannigan et al. [2020] discovered compelling evidence in Milgram’s personal archive that lends weight to Orne & Holland and Mixon’s point of agreement. They first note:

In his first journal article about his obedience research, Milgram [...] stressed the dramaturgical credibility of the experiment. He emphasized that “[w]ith few exceptions subjects were convinced of the reality of the experimental situation” [...]. The implication was that the subjects fully believed that what was happening was real, and despite indications that the learner was in increasing pain, 26 out of 40 proceeded to administer the maximum shock. [Perry, Brannigan et al. 2020, 92]

However, in conflict with this, Perry, Brannigan et al. add that what Milgram failed to mention was that in 1962 he instructed his research assistant Takeo Murata to analyse the subjects’ post-experimental interview survey data and that he do so with a specific focus on the issue of subject deception. The report Murata wrote compared:

[...] the degrees of obedience between those subjects who said they were doubtful that the shocks were painful and those who were certain they were. [...] Murata [1962:1] wrote, “The following is a condition-by-condition analysis to determine whether shock level reached was affected by the extent to which the subject believed that the learner was actually receiving shock.” [...] The report’s main results were a comparison of the mean shock levels administered by subjects who had “fully believed” [...] in the reality of shocks versus those classified as having “not fully believed” [...]. Murata found that in 18 of 23 experiments, those subjects who fully believed the learner was getting painful shocks gave lower levels of shock than subjects who doubted the shocks were real. [Perry, Brannigan et al. 2020, 94]
The reverse also applies: “Those who were less successfully convinced by the cover story were more obedient” [2020, 99]. Put differently: “obedience increases with skepticism of pain” [2020, 98], a conclusion that, across most of Milgram’s experiments, bolsters Orne & Holland’s and Mixon’s point of agreement that Milgram failed to sufficiently deceive his subjects into believing the shocks were dangerous.

Perry, Brannigan et al. further reinforced their point by referencing Hollander & Turowetz’s analysis of obedient subjects’ post-experimental justifications for completing. That is, across the five conditions for which, at Milgram’s personal archive, audio-recordings were available, Hollander & Turowetz [2017, 661–662] found the most common reason 46 obedient subjects provided for inflicting every shock was because they did not think the learner was being harmed (33 subjects, or 72 percent).

So what emerges across Orne & Holland, Mixon, and Perry, Brannigan et al.’s research is a common denominator: Milgram’s baseline procedure failed to sufficiently deceive the “obedient” subjects into believing the shocks were harmful and, as a result, these subjects developed a belief that no harm would come from their compliant actions. And this mono-causal account explains why most New Baseline subjects completed the experiment. Although all three accounts are technically methodological critiques of the Obedience Studies, all provide dismissive explanatory accounts for Milgram’s apparently counter-intuitive results.

However, the validity of Orne & Holland, Mixon, and Perry et al.’s arguments have been challenged [see Russell 2018, 120–126], [Russell & Gregory 2021, Russell 2024]. The crux of this counter-critique is that largely because of the partition separating the subjects from the learner, subjects in the New Baseline could not know for certain if the shocks really were harmful. The wall introduced a strong dose of ambiguity into the New Baseline procedure. Unable to see the learner, “many” subjects were left in, as Perry notes, “a state of uncertainty and stress” [2013, 162]. This uncertainty meant any decision by a skeptical subject to inflict every shock necessitated they take a major risk: their suspicion that the experiment was possibly, probably, or even most certainly harmless could be wrong and, if so, the learner would be seriously injured. For subjects who took this risk, an important question emerges: was it morally correct for them to prioritise their mere suspicion the experiment was fake and inflict every shock because they thought everything would be “all right”?1

---

1. As mentioned, Mixon argued that Milgram’s purposeful insertion of ambiguity left subjects with little choice but to side with the only expert present: the experimenter. Mixon therefore suspected that if Milgram eliminated—instead of purposefully inserted—ambiguity into the basic procedure, subjects would have known if the shocks really were harmful and acted accordingly (if the shocks were real, they would not have completed the experiment). We disagree. Rather, the ambiguity Milgram injected into the basic procedure introduced the possibility that the obedient subjects’ hunch that the learner was not seriously being harmed could
There is, it transpires, a morally correct and inherently safe resolution to the dilemma over whether to stop or continue inflicting shocks: if a subject was, as all must have been, unsure if the shocks were real, the safest (moral) choice was to err on the side of caution and prematurely end the experiment [Coutts 1977, 520 cited in Russell 2018, 125]. Doing so eliminated all risk of a subject’s hunch being wrong and instead guaranteed the safety of a fellow human being [Russell & Gregory 2021]. Although this ethically cautious type of problem-solving was uncommon, a minority of subjects exhibited it across numerous experimental conditions. For example, one suspicious and uncooperative subject later said, “When I decided that I wouldn’t go along with any more shocks, my feeling was ‘plant or not [...] I was not going to take a chance that our learner would get hurt’ ” [Russell 2018, 124]. Another subject in a different condition noted he “wasn’t sure” if the learner “was getting the shocks,” but when “he started to complain vigorously [...] I refused to go on” [Russell 2018, 142]. In fact, one subject explicitly stated he was certain the subject was not being shocked, but he refused to trust the experimenter:

**TEACHER:** “I don’t believe this! […]”  
**EXPERIMENTER:** “You don’t believe what? […]”  
**TEACHER:** “I don’t believe you were giving him the shock.”  
**EXPERIMENTER:** “Then why, why won’t you continue?”  
**TEACHER:** “Well I, I just don’t want to take a chance […]” [Russell 2018, 124]

This subject did not believe the shocks were real but, largely because of the wall separating him from the learner, he could not be certain. The subject’s uncertainty in the ambiguous situation confronting him dictated that he could not afford to take that chance and—contra Mixon—trust the expert in charge because there was still a possibility his hunch might be wrong. And this subject was obviously aware of the consequences that such a mistake would have on the learner.

But in the *New Baseline* condition most subjects resolved the experiment’s inherent moral dilemma differently: despite the risk that the shocks might have been real, 65 percent of them assumed that no harm would come from their completing the experiment and that doing so would probably be “all right.” Of course, because all *New Baseline* subjects remained uncertain over the reality of the shocks, deciding to risk another person’s well-being over a mere hunch was unethical (especially when compared to the option of prematurely stopping what might have been a dangerous experiment).

have been wrong, which, as mentioned, was a possibility that came with potentially devastating consequences for the learner. If all subjects could have been certain that the learner was not being seriously harmed, there would have been no chance of them being wrong. And if there was no possibility of being wrong, then subjects would not have been faced with resolving a moral dilemma, which is what the experiment ultimately tested. In sum, no ambiguity, no moral dilemma.
Thus, the mystery at hand returns: why did 65 percent of New Baseline subjects choose to act so unethically? Perhaps the most discussed explanatory account of the Obedience Studies was that eventually offered by Milgram [1974]; probably because of the widely shared assumption that surely the inventor of this experiment knew why most completed the baseline.

4 The Agentic State

In 1974 Milgram published his book *Obedience to Authority: An Experimental View*. To give this book an air of completeness, Milgram greatly desired it include a mono- over a multi-causal theoretical account. Thus, from the late 1960s he started developing his best attempt at producing such a theory: his so-called *Agentic State*. So, what did this mono-causal theory entail?

In his book Milgram, drawing on evolutionary theory, argued that, like many animals, humans function in hierarchical structures. Due to the survival value of organization over disorganization, he argued hierarchical structures have, through breeding, led humans to gain advantages over their competitors in the natural environment. Milgram is careful to point out that obedience to those higher in the chain of command is not innate. “Rather, we are born with a potential for obedience, which then interacts with the influence of society to produce the obedient man” [1974, 125]. He then argued that mankind’s potentially destructive impulses are frequently kept in check by the conscience of an autonomously acting individual. However, when an individual is introduced into a hierarchical chain of command, a homeostatic-driven internal change can take hold in which conscience-driven control over their actions can be suppressed and supplanted by a higher authority’s demands. Milgram termed this process the *Agentic Shift*, which is where “[...] a man [sic] feels responsible to the authority directing him but feels no responsibility for the content of the actions that the authority prescribes” [1974, 145–146].

After this shift, the previously autonomous individual enters the *Agentic State*, which is where “the individual no longer views himself as responsible for his [sic] own actions but defines himself [sic] as an instrument for carrying out the wishes of others” [1974, 134]. In this state subjects became so-called exemplary functionaries: solely concerned with—to the full satisfaction of the experimenter—how effectively they carried out their instructions.

Milgram believed two main factors put subjects into the *Agentic State*: his previously mentioned Strain Resolving Mechanisms and Binding Factors. In explaining both compliance and defiance, Milgram [1974, 154] succinctly captured his theory as follows:

\[ O; B > (s - r) \]
\[ D; B < (s - r) \]

in which \( O \) represents obedience; \( D \), disobedience; \( B \), binding factors; \( s \), strain; and \( r \), the strain-resolving mechanisms. Obedience
is the outcome when the binding factors are greater than the net strain (strain as reduced by the resolving mechanisms), whereas disobedience results when net strain exceeds the strength of the binding forces.

Considering Milgram’s earlier journey of discovery, one can see why he found aspects of this post hoc theory appealing. That is, the Agentic State’s above concise mathematical-like formula is not only mono-over multi-causal [Gibson 2019, 61], but it is also a theory of social engineering that closely aligns with his invention of the Obedience Studies: the more manipulative forces he added, the more efficacious his basic procedure seemed to become.

That said, Milgram’s theory has attracted much criticism [see Russell 2018, 132]. One telling critique is that the Agentic State cannot explain the variance in completion rates across the different experimental conditions [Blass 2004, 216]. Neither can it account for individual differences within conditions [Gibson 2019, 56–57]. So although the Agentic State attempted to offer a comprehensive mono-causal or “totalizing” explanation that illuminated “behaviour across individuals, across experimental conditions, and across time” [2019, 61], it failed to deliver.

Even scholars noted for being positively disposed toward Milgram’s research admit there is “very little empirical support” for the Agentic State [Miller 2004, 210] and that the book’s “weakest” chapter is that which presents his theory [Blass 2004, 216]. Again, after Milgram had been trapped for approaching a decade in a mono-causal theoretical drought, the Agentic State was likely his last-ditch attempt at offering such an explanation. It transpires that the creator of the experiments did not know why subjects completed his basic procedure.

Although in 2004, Obedience-study specialist Arthur Miller argued, after the demise of the Agentic State, that there is “no conclusive theory to account for destructive obedience—or defiance, either” [2004, 233], a decade later he alluded to a “[...] most promising recent development [...]” [2004, 568]. Miller was referring to Engaged Followership theory.

5 Engaged Followership

The leading authors of Engaged Followership theory—Haslam & Reicher—argue that subjects in the Obedience Studies:

[...] are able to inflict harm on others not because they are unaware that they are doing wrong, but rather because—as engaged followers—they know full well what they are doing and believe it to be right [...] [Haslam, Reicher et al. 2015a, 79]
This engagement, they add, involved subjects identifying with a “noble” cause [2015a, 79], specifically the subjects’ personal affiliation with the experimenter’s “scientific mission”: the research’s purported quest to explore the effects of punishment on learning [2015a, 77]. Thus, the experimenter’s explicit and implied appeals to his scientific mission—and the greater good so frequently associated with it—persuaded then motivated most baseline subjects to endure the victim’s pained reactions until every shock had been inflicted. As the authors themselves conclude:

[...] it appears that the behavior of Milgram’s participants was underpinned by active commitment to his [scientific] endeavor and an associated desire to prove themselves as worthy subjects. Indeed, it was this commitment and desire that appear to have steeled them to persist in the face of the significant stressors that the task exposed them to. [Haslam & Reicher 2017, 71]

As the pseudonymous subject Morris Braverman worded it: “[...] in the interest of science, one goes through with it” [Milgram 1974, 54].

Haslam & Reicher have published a variety of empirical studies that they argue provides evidence in favour of their theory [Reicher, Haslam et al. 2012, Haslam, Reicher et al. 2014, 2015b,a].

According to Gibson, the “crucial variable” to emerge from this research is that of “social identification” and, as a result, Haslam & Reicher found:

[...] conditions that promote identification with the experimenter lead to higher levels of completion; conditions that promote higher levels of identification with the learner lead to higher levels of defiance. [Gibson 2019, 59]

If true, Engaged Followership theory purports to explain not only why subjects completed the New Baseline, but also why they resisted:

[...] being torn between two insistent voices—the one representing science, the other representing the community of ordinary people—the question of which one the participant heeds ultimately depends on which category he identifies with most. [Reicher, Haslam et al. 2012, 319]

This intriguing theory, however, has attracted some criticism. For example, Gibson [2019, 60] notes, despite the above evidence in support of their account, Haslam & Reicher’s theory has never directly been tested on an Obedience Study-type procedure, adding:

[...] it is notable that the only study that involves direct measures of naïve participants’ levels of identification in a Milgram-esque paradigm [Haslam et al., 2014] found that there were no effects
of identification on participants’ likelihood of completing the experiment, likelihood of completing post-experiment measures, or on the length of time that participants continued with the experiment. This raises the question of how well the somewhat-limited evidence available to date actually supports this account of Milgram’s findings.

Other critics like Hollander & Turowetz note that during Milgram’s post-experimental interviews, subjects frequently stated they completed the experiment for a range of reasons that had nothing to do with advancing the experiment’s scientific aims. More specifically, and as touched on earlier, Hollander & Turowetz note that out of their sample of 46 obedient subjects, most provided the following (sometimes multiple) accounts for completing: because they did not think the learner was being harmed: 33 (72 percent), because they were following instructions: 27 (59 percent), or because they felt contractually obligated into completing: 5 (11 percent). Subjects only mentioned they inflicted every shock because of the importance of the experiment 11 times (24 percent) [Hollander & Turowetz 2017, 661–662]. Hollander & Turowetz thus suspect there are likely “multiple empirically grounded”, thus a multi-causal array of “explanations” of the Obedience Studies [Hollander & Turowetz 2018, 302].

Because Hollander & Turowetz found subjects offered a variety of different justifications for completing, they accused Haslam & Reicher of advancing a mono-causal explanatory account that is simply not true [2017, 671–672], [see also Gibson 2019, 61]. Haslam & Reicher [2018, 293] responded:

In proposing the “engaged followership” account our intention was never to try to replace one monolithic orthodoxy [Milgram’s Agentic State] with another and to try to have the final word. We entirely endorse H[ollander] & T[urowetz]’s argument that it would be both futile and wrong to posit a single process to account for what Milgram found [...].

Thus, here, on top of now agreeing that there are probably a variety of possible reasons subjects inflict every shock, Haslam & Reicher willingly

---

2. An issue with Hollander & Turowetz’s thesis is if most obedient subjects, as they assert, completed Milgram’s basic procedure because they did not think the learner was being harmed, they are effectively siding with Orne & Holland, Mixon, and Perry et al.’s point of agreement: Milgram’s attempts to deceive his subjects failed. Therefore, it can be argued that Hollander & Turowetz are, by implication, supporters of the Incredulity Hypothesis, and thus the survival of their thesis—recently converted into a book—will be tied to its fate.

3. Although they may not have intended for their theory to be mono-causal, it does closely resemble such an account: “Our reconceptualization also has the advantage of invoking a single set of processes to explain not only when people follow authority but also when they resist it” [Reicher, Haslam et al. 2012, 322].
concede that *Engaged Followership* theory can only account for why some, and not all, subjects completed the Obedience Studies.

That said, Haslam & Reicher have countered aspects of Hollander & Turowetz’s critique. For example, Haslam & Reicher used *Engaged Followership* theory to account for Hollander & Turowetz’s finding that most obedient subjects in their sample later justified completing because they did not think the learner was being harmed:

Participants are induced to cooperate because they accept the Experimenter’s assurances about the study which in turn is because they accept the credibility of his statements as a scientist. [Haslam & Reicher 2018, 294]

They then, in different words, repeat this point:

More concretely, to say “I imposed painful shocks against the learner’s will because I think science is a good thing” seems contradictory unless one goes on to explain that because one believes in science one also trusts that a legitimate scientific experiment will not harm people. [Haslam & Reicher 2018, 295]

The problem, however, with Haslam & Reicher’s counter argument is, in an attempt to defend their embattled theory, they deploy one of Mixon’s foundational arguments from his methodological critique of the Obedience Studies: because “large chunks” of the modern world “can be understood only by experts”, ordinary people in confusing situations place their trust in experts who know best: the experimenter said the seemingly painful shocks were not dangerous, so obedient subjects knew shock-infliction was harmless. Therefore, according to Mixon, most subjects completed the *New Baseline* not because, as Milgram argued, they buckled under the coercive power of malevolent authority, but because they knew it was safe to do so. Of course, if Mixon’s methodological critique is correct (and Haslam & Reicher’s above quotes inadvertently implies it is), then what is the use in them trying to provide an explanation for the results of a study that has such a flimsy methodological foundation? Put differently, Haslam & Reicher’s pro-Mixon counter argument reinvigorates the ability of their theory to explain the results of a social science experiment that—according to Mixon—is methodologically flawed.

*Engaged Followership* theory is not immune to another perhaps even more damning criticism. Haslam & Reicher cannot evade the following fact: among all those subjects who later justified that they completed the experiment to help advance “science,” some of this group also simultaneously tried to undermine (sabotage?) the experiment’s purported scientific goals [Russell 2018, 230–231]. For example, consider the pseudonymous and fully “obedient” Eleonor Rosenblum who, during her post-experimental debrief, justified completing because: “It is an experiment [...]. So I had to do it”
[Milgram 1974, 83]; later also declaring “I’m one for science [...]” [1974, 83]. So, although Haslam & Reicher now concede that not every subject offered this justification for completing, here is a clear case where Engaged Followership theory seems to rather precisely account for why a subject inflicted every shock. As a supporter of science, Rosenblum believed that inflicting the shocks was, as Haslam & Reicher would put it, “right [...]” [Haslam, Reicher et al. 2015a, 79].

However, when Rosenblum was later reintroduced to the unharmed learner she confided to him, in a statement whose accuracy was corroborated by the measured recording of her performance, “‘Did you hear me stressing the [correct] word[?] I was hoping that you would hear me’” [Milgram 1974, 82]. That is, Rosenblum covertly attempted to provide the learner with the correct response to her questions by vocally emphasizing the correct answer. Here Rosenblum engaged in what Milgram termed “subterfuge” [1974, 159], because she hoped her provision of the right word would enable the learner to correctly answer her questions thereby enabling her to avoid shocking him. Of course, an obedient subject who genuinely identified with the experiment’s scientific mission and believed shock infliction to be “right”, should not have covertly tried to undermine the very purpose that they (apparently) so identified. So, although Rosenblum’s behaviour appears to be a textbook case in support of Haslam & Reicher’s theory—she said she inflicted the shocks it for science—it appears their wilting theoretical account cannot even explain her behaviour. What further detracts from the potency of Haslam & Reicher’s theory is that Rosenblum was no outlier. As Milgram observed:

Some subjects could be observed signaling the correct answer to the victim by stressing it vocally as they read the multiple-choice words aloud. That is, they attempted to prompt the learner and thus prevent his receiving shocks. [Milgram 1974, 159]

And Milgram’s detection of acts of subterfuge have independently been confirmed by others:

Hannah wasn’t the only subject who defied Williams [the experimenter] by accentuating the right answer—I’d heard it on a number of recordings. [...] another subject said: “I did everything in my power to emphasise the correct answers [...] and I hoped the supervisor watching it didn’t catch on [...].” [Perry 2013, 124]

There was even some diversity in these subjects’ techniques of subterfuge: some gave the appearance they were inflicting intensifying shocks, but instead covertly attempted to avoid fully depressing the shock switches [Perry 2013, 197], [Gibson 2019, 183]. Subterfuge occurred so frequently that Milgram had to instruct the experimenter to ensure subjects performed their set tasks correctly, with him telling one errant subject: “Please read the words in an even tone” [Perry 2013, 196]. This subject then attempted to lightly brush
the shock switches, to which the experimenter responded: “That didn’t make contact. Press the switch down completely” [2013, 197]. Nonetheless, the key point being is that Engaged Followership theory struggles to account for subterfuge-type behaviours.

Although Haslam & Reicher’s theory exhibits various weaknesses, it still offers an important piece of the theoretical puzzle; remembering that when Milgram was inventing the experiments he would not have gotten far without the legitimating justification of harm-infliction for science. As Burger [2009, 3] argued:

Milgram’s experimenter was granted the legitimacy of authority by virtue of his association with the experiment, the university, and perhaps even science.

However, we suspect the aura of science is less explanatory and more of a singular and no doubt important Milgram-like Strain Resolving Mechanism: any means that, as mentioned, helped subjects offset the tension typically associated with harm infliction. And the term for this specific Strain Resolving Mechanism is Moral Inversion. Moral Inversion is where harm infliction is persuasively transformed from “something evil” (shocking an innocent person) into something “good” (advancing human learning) [Adams & Balfour 1998, xx, cited in Russell 2018, 59].

For us, the above criticisms against Engaged Followership theory sees its relegation from a mono-causal, to a partial, to—at best—a minor and incomplete explanation for Milgram’s results (again, better subsumed under the overriding banner of Moral Inversion). We agree with Hollander & Turowetz that “[...] the puzzle of Milgramesque behaviour is more complex than engaged followership at present allows” [2017, 672].

In conclusion of this section, a critical assessment of what are arguably the three most discussed explanatory accounts as to why ordinary people completed Milgram’s New Baseline experiment has been provided. We dismiss all three accounts, largely because, in a variety of ways, their assertions are inconsistent with Milgram’s findings. To avoid leaving the reader at an explanatory dead-end, we also believe less prominent theories should be granted greater critical attention. The next section presents one such theory.

---

4. Haslam & Reicher could respond to this point by noting that, as they say, “[...] even when participants follow the experimenter, there is still a pull toward the learner—a concern with his fate. [...] To some extent at least, participants attend to both voices” [Reicher, Haslam et al. 2012, 323]. Our issue with this response is with what they say next: “Identification determines which is accorded most weight” [2012, 323]. It is unclear who Rosenblum accorded the most weight—had the experiment really been exploring punishment on learning, the experimenter would have obtained tainted ‘data’ and she also ignored the fettered learner’s desperate requests to be freed. For reasons discussed in the Footnote 6, the person we suspect Rosenblum accords the most weight (“identified” most with) was herself: her needs and desires.
6 Fear of a confrontation

As already touched on, one potentially promising theory is inherent within Milgram’s earlier invention of his procedure: as Milgram, in the role of a social engineer, added more strain resolving mechanisms and binding factors to his basic procedure, the cumulatively more coercive the experiment became. This explanation may help explain how Milgram ensured most subjects completed the baseline, but it does not account for why exactly they chose to do so. Thus, the motivational impetus remains a mystery. In his book Milgram proposed, beyond his Agentic State, another reason why some subjects may have decided to complete. He observed:

If sympathetic concern for the victim were the exclusive force, all subjects would have calmly defied the experimenter. Instead, there were both obedient and defiant outcomes, frequently accompanied by extreme tension. [...] Therefore there must be a competing [...] inhibition that precludes activation of the disobedient response. [Milgram 1974, 42-43]

If the competing inhibition could be identified, the base motive/s would be revealed. One possible competing inhibition Milgram mentioned that may have precluded disobedience was:

The learner cannot break off and at the same time protect the authority’s definitions of his own competence. Thus, the subject fears that if he breaks off, he will appear arrogant, untoward, and rude. Such emotions [...] suffuse the mind and feelings of the subject, who is miserable at the prospect of having to repudiate the authority to his face. [Milgram 1974, 150]

Others have taken this potential explanation for completing further with what, it will be noted, is a multi-causal account: “a socially awkward confrontation with the experimenter was actually the most common—but not the only—[...] motivational force behind the New Baseline experiment’s high completion rate” [Russell 2018, 201], [see also Russell & Gregory 2011]. For the reader, this probably sounds like too trivial a reason for completing, but Milgram’s results lend much weight in favour of it: in nearly all experiments where subjects, in order to prematurely end the experiment, did not have to initiate and then engage in a direct confrontation with the experimenter, none,

5. This explanation is multi-causal because if most subjects completed for this reason, a minority of subjects competed for another or even other quite different reasons [see Russell 2009, 127]. Although this explanation is multi-causal, instead of presenting a wide and potentially confusing array of potential factors or mechanisms (like [Milgram 1963], [Burger 2009]), in line with the advantage with the mono-causal accounts, the above explanation retains elements of exactitude and simplicity: it asserts why most subject complete.
or few of them inflicted every shock (average completion rate: 11.25 percent). This account suggests that in the absence of having to engage in a direct confrontation to stop the experiment, nothing or little competed with most subjects’ sympathy for the learner’s predicament, leaving them free to act on their personal preference to stop inflicting further shocks. However, in nearly all the experiments that necessitated subjects, in order to end them prematurely, initiate then engage in a direct confrontation with the experimenter, relatively large percentages of subjects inflicted every shock (average completion rate: 54.61 percent (see Table 1 below)).

A statistical test based on Table 1 justifies the conclusion that the hypothesis that a fear of confrontation is causally relevant for “obedient” behaviour should be accepted (Chi Square obedient vs. disobedient in the two kinds of conditions, 1d.f. = 123; p < 0.01).

Table 1: Experimental conditions where siding with the learner did or did not require the subject to initiate then engage in a confrontation with the experimenter [Russell & Gregory 2011], [Russell 2018, 79–80; 92–96; 202].

<table>
<thead>
<tr>
<th>Experiment (in order of lowest to highest completion rates)</th>
<th>Nb of subjects</th>
<th>Nb of subjects who completed (were “obedient”)</th>
<th>% completion rate (“obedience” rate)</th>
<th>To stop the experiment, did subjects have to initiate then engage in a confrontation with the experimenter?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learner Demands to be Shocked</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>No</td>
</tr>
<tr>
<td>Authority as Victim</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>No</td>
</tr>
<tr>
<td>Two Authorities – Contradictory Demands</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>No</td>
</tr>
<tr>
<td>Subject Chooses Shock Level</td>
<td>40</td>
<td>1</td>
<td>2.5</td>
<td>No</td>
</tr>
<tr>
<td>Groups for Disobedience</td>
<td>40</td>
<td>4</td>
<td>10</td>
<td>No</td>
</tr>
<tr>
<td>Relationship</td>
<td>20</td>
<td>3</td>
<td>15</td>
<td>Yes</td>
</tr>
<tr>
<td>Group Pressure and Action Against a Person</td>
<td>40</td>
<td>7</td>
<td>17.5</td>
<td>No</td>
</tr>
<tr>
<td>Experimenter Absent</td>
<td>40</td>
<td>9</td>
<td>22.5</td>
<td>No</td>
</tr>
</tbody>
</table>

6. We suspect the previously discussed Eleonor Rosenblum was a subject whose behaviour is explicable by the fear that, in order to help the learner by stopping the experiment, she would have to initiate then engage in a direct confrontation with the experimenter. In fear of this solution, she then sought out, regarding the moral dilemma before her, an alternative and more attractive course of action. That is, she came to suspect that if she vocally emphasized the correct answer to the learner, although she would technically taint the ‘accuracy’ of the scientist’s data, doing so would enable her to avoid having to shock the learner (which would thereby obviate any learner-generated pressure that she help them by initiating then engaging in a direct confrontation with the experimenter).
Table 1 excludes the Ordinary Man Gives Orders (4/20 or 20 percent completion rate) and Teacher In Charge (11/20 or 55 percent completion rate) conditions because both experiments necessitated the subject engage in a confrontation not with the experimenter (who was absent and, unlike during the Experimenter Absent condition, could not be contacted by phone), but a confederate posing as a second teacher.

In this condition the experimenter gave his usual instructions, but then stated he had to leave the laboratory and instructed the subject to complete the experiment on his own. The subject was given the experimenter’s contact telephone number. But if, in concern for the learner, the subject rang the experimenter, the experimenter would urge them to continue inflicting further intensifying shocks. Interestingly, sometimes a subterfuge-type behaviour emerged: “Though continuing with the experiment, several subjects administered lower shocks than were required and never informed the experimenter of their deviation from the correct procedure” [Milgram 1965, 66]. Other subjects told the experimenter over the phone that they were inflicting intensifying shocks, but only inflicted low-intensity shocks. All of these subjects, it seemed, “found it easier to handle the conflict in this manner than to precipitate an open break with authority” [1965, 66]. In support of the central role of a fear of a confrontation, if a subject “refused to give higher shocks when commanded by telephone”, the experimenter could return and “frequently force further obedience [...]” [1965, 66].

In this condition the subject arrived at the laboratory to find a note informing them that the experimenter was unable to attend and that they were to press a nearby cassette player’s “play” button (which relayed the experimenter’s usual instructions). Although in this condition the subject, in order to stop the experiment, did not have to initiate then engage in a confrontation with the experimenter (there was nobody to confront), it still generated a relatively high 37.5 percent completion rate. Then again, due to “technical difficulties” Milgram did not seem to have much methodological faith in this experiment (he clearly did not believe it merited inclusion in his 1974 book and few published details exist on it) [1965, 65–66].

The most obvious exception to this general rule was the Relationship condition, where subjects were instructed to bring another person along to the laboratory who was at least an acquaintance they had known for two or more years (3 of the 20 subjects, however, brought a relative). On
arrival, one of the pair became the teacher and the other the learner (then after the learner was strapped into the electric chair and the teacher and experimenter left the “shock” room, Milgram appeared before the learner and explained the experiment’s actual purpose: to test if their friend would shock them. Milgram then quickly trained these learners how to react to the intensifying “shocks.” But in conflict with Mixon’s assertion that people in ambiguous situations tend to side with experts, 85 percent of subjects in this condition refused to do so [Russell 2014]. What the Relationship condition illustrated is that a teacher’s fear of having to confront the experimenter in order to stop the experiment diminished greatly when the victim was someone they knew. In fact, this fear appeared to completely evaporate when the acquaintance happened to be a family member: 0 out of 3 (0 percent) completed. This result—despite being Milgram’s greatest ethical breech—hints at an important insight that, we would argue, alludes to the inner-workings of the basic procedure: excluding the Relationship condition, all of Milgram’s baselines and experimental variations involved a stranger as the victim. And with stranger victims, it was obviously, relatively speaking, easier for Milgram to coerce (socially engineer) many people with his barrage of binding factors and strain resolving mechanisms into completing (especially when compared to him pushing subjects into hurting a person who was an acquaintance and, especially, a family member). In sum, factors like victim relationship in conjunction with Milgram’s mention of the fear of a confrontation—and, not discussed here, the incomparably powerful strain-resolving power of the shock generator [Russell 2018, 239–245]—are all potential insights that we believe require greater critical attention. See Russell [2018] for a fuller account of this theory.

7 Conclusion

This article has presented and critically assessed what are arguably the three most discussed and thus prominent explanatory accounts as to why ordinary people completed Milgram’s New Baseline experiment. The theories reviewed included the Incredulity Hypothesis, the Agentic State, and Engaged Followership. We dismissed all three theories because, in a variety of ways, their assertions are inconsistent with what Milgram found. Consequently, we believe less prominent theories exploring subject motivation—like the subjects’ fear of having to engage in a confrontation with the experimenter—requires closer critical attention.

But despite our assertions on what theories deserve less or more attention, what should become apparent to the reader is that in relation to what is perhaps both the most revered and reviled experiment in social psychology—a study that for well over half a century has, across both scholarly circles and in popular culture, sustained enormous attention—the gains on the theoretical
frontlines have been miniscule. But the positive side of this theoretical deficit is that for budding social scientists, this likely means that the mystery of why subjects completed the baseline experiment remains.

Acknowledgements

The authors would like to thank Emeritus Professor Robert Gregory of Victoria University of Wellington, New Zealand, for his comments. All responsibility rests with the authors.

Bibliography


Doris, M., John, Niemi, Laura, et al. [2024], The incredulity hypothesis and the enduring legacy of the obedience experiments, Philosophia Scientiæ, 28(2), 53–89.


—— [2015a], “Happy to have been of service”: The Yale archive as a window into the engaged followership of participants in Milgram’s ‘obedience’ experiments, *British Journal of Social Psychology*, 54(1), 55–83, doi: 10.1111/bjso.12074.


RUSSELL, Daniel C. [2009], Practical Intelligence and the Virtues, Oxford: Oxford University Press.


—— [2024], Stanley Milgram’s obedience studies: A contemporary ethical and methodological assessment, Philosophia Scientiae, 28(2), 107–129.


SLATER, Lauren [2004], Opening Skinner’s Box: Great Psychological Experiments of the Twentieth Century, New York: W. W. Norton.