On the Automaticity and Ethics of Belief

Uwe Peters
Centre for Logic and Philosophy of Science, KU Leuven
Department of Economics, UCL

[Published in 2017; invited contribution to *Ethics, Law, and Cognitive Science*; special edition of *Teoria*, edited by Mario De Caro and Massimo Marraffa, 99–115.]

Abstract

Recently, philosophers have appealed to empirical studies to argue that whenever we think that *p*, we automatically believe that *p* (Millikan 2004; Mandelbaum 2014; Levy and Mandelbaum 2014). Levy and Mandelbaum (2014) have gone further and claimed that the automaticity of believing has implications for the ethics of belief in that it creates epistemic obligations for those who know about their automatic belief acquisition. I use theoretical considerations and psychological findings to raise doubts about the empirical case for the view that we automatically believe what we think. Furthermore, I contend that even if we set these doubts aside, Levy and Mandelbaum's argument to the effect that the automaticity of believing creates epistemic obligations is not fully convincing.

Introduction

It is widely accepted that we are able to think about or entertain propositions without believing them (Descartes 1644/1984; Fodor 1983; Dennett 1987; McDowell 1998; O'Brien 2007; McHugh 2011; Kriegel 2013). However, some philosophers have employed cognitive scientific findings to argue that this view is in fact false (Millikan 2004; Mandelbaum 2014; Levy and Mandelbaum 2014).

For instance, Millikan (2004) holds that there are psychological studies providing "evidence that when we hear someone speak, normally what is said goes directly into belief [...]. We do not first understand what is said and then evaluate whether to believe it" but rather immediately accept[1] the information that we are presented with (121). Similarly, on the basis of empirical research, Mandelbaum (2014) maintains that merely "thinking" that *p* "is believing" that *p* (55).

The claim is that whenever we entertain a proposition *p*, we will automatically and prior to analyzing the truth of *p* come to believe *p* at the unconscious level. Upon subsequent reflection at the conscious level we may reject *p* or endorse[2] *p* but that is only possible after we have initially accepted it at the unconscious level. This view of belief formation is often called the *Spinozan theory*, as Spinoza (1677/1982) is thought to be the first who defended it.[3]

Some empirically oriented philosophers who advocate the Spinozan theory hold that the theory has implications for the ethics of belief. For instance, after arguing for the Spinozan theory, Levy and Mandelbaum (2014) continue that people "who know about" their "propensities" to believe propositions

[1] There are differences between accepting or affirming that *p* and believing that *p* (e.g., one may for the sake of the argument accept *p* without believing *p*). However, for the purpose of this paper, I shall ignore this point and follow Spinozans in treating accepting or affirming that *p* as initiations of believing that *p* (Mandelbaum 2014: 61) or at any rate as leading to a
[2] Spinozans use the terms 'to endorse *p*' (Mandelbaum 2014) or 'to certify *p*' (Gilbert 1991) to distinguish conscious, subject-controlled affirmations of *p* from unconscious, automatic affirmations of *p*, for which they tend to use the term 'to accept *p*'.
[3] Spinoza did not use the conscious vs. unconscious distinction that contemporary Spinozans invoke in their account of belief formation.

through merely entertaining them have epistemic "obligations to take the risk of forming unjustified" and "immoral beliefs into account" when they expose themselves to them (30).

In the following, I use theoretical considerations and data from psychological studies to cast doubts on the empirical case for the view that we automatically believe the propositions that we entertain. In addition, I maintain that even if we set these doubts aside, Levy and Mandelbaum's argument to the effect that the automaticity of believing creates epistemic obligations is not fully convincing.

1. The Spinozan theory

The most developed form of the Spinozan theory, which is also the version that I will focus on here, has been introduced by Gilbert and his colleagues (1990, 1991, 1993) and elaborated by Mandelbaum (2014). It can be summarised in the following three claims.[4]

> (1) People do not have the ability to contemplate propositions that arise in the mind […] before believing them. Because of our mental architecture, it is (nomologically) impossible for one to not immediately believe propositions that one tokens.
>
> (2) Accepting a proposition is accomplished by a different system than rejecting a proposition. Because different systems are at play, the processes of accepting and rejecting should be affected by performance constraints in different ways. […]
>
> (3) Forming a belief is a passive endeavour.[5] However, rejecting a proposition is an active and effortful mental action, which can only happen after a belief has been acquired. Consequently, one can effortlessly form new beliefs while being mentally taxed, but rejecting an already held belief will become more difficult the more mentally taxed one is. (Mandelbaum 2014: 61)

Based on these claims, the Spinozan theory yields a number of predictions. For instance, when a subject is presented with a proposition *p* and prevented from rejecting *p* (for instance, by being distracted), she should not remain doxastically neutral about *p* but end up believing the proposition. Furthermore, since, according to the Spinozan theory, it is "(nomologically) impossible for one to not immediately believe propositions that one tokens" (ibid), subjects should have the tendency to accept *p* even when they are *before* they are presented with the proposition told that *p* is false.

Spinozans have appealed to empirical studies to argue that these predictions are borne out by the data. I'll briefly consider some central examples.

2. Evidence supporting the Spinozan theory

Among the psychological work that Spinozans heavily rely on are the following two experiments conducted by Gilbert and colleagues (Millikan 2004; Mandelbaum 2014; Levy and Mandelbaum 2014).

[4] Mandelbaum (2014) adopts a fourth, negation-related claim that I shall not consider here.

[5] It is worth noting that even though advocates of the Spinozan theory claim things such as (i) in general "[f]orming a belief is a passive endeavour" (Mandelbaum 2014: 62) or (ii) "we, strictly speaking, do not form beliefs for reasons at all" (Levy and Mandelbaum 2014: 17), these claims are – even on the Spinozan view itself – not correct. For, according to the Spinozan theory, one can, once one has automatically accepted *p,* still reject, or endorse *p* at the conscious level, and this will at least sometimes happen for epistemic reasons pertaining to whether *p* is true. Furthermore, the subsequent rejection or endorsement of *p* will result in a belief itself, namely in the belief that *p* is false or true, respectively. Since rejecting and endorsing *p* are on the Spinozan account "active" (Gilbert 1991: 108; Gilbert et al. 1993: 4; Mandelbaum 2014: 61), it follows that some cases of belief-formation are, against claims (i) and (ii), even according to the Spinozan theory active in nature and based on reasons (e.g., cases of rejection-, or endorsement-based belief-formation).

Gilbert et al. (1990) asked subjects to learn statements about the meaning of words in a fictional language, for instance, 'A *monisha* is a star'. Each statement appeared briefly on a screen and was followed by a validating term, i.e., 'true' or 'false'. On some trials, during the learning phase, subjects had to identify musical tones that rang out after the validating word appeared. This was meant to drain their mental resources. In the testing phase, subjects were then again shown the sentences and asked whether they were true, false, or not present during the learning phase.

Gilbert et al. found that participants who were distracted during the validation process by the tone-identification task didn't manage to remain doxastically neutral about the statements presented to them but tended to encode the sentences, including those marked as false, as true. Gilbert et al. and others Spinozans take this to show that we "first believe what is said and then, if we are not under too much cognitive stress, we may think it over critically and reject it" (Millikan 2004: 121).

Gilbert et al. (1993) conducted another study that led to similar findings. Subjects were asked to read two crime reports that included both true and false statements. True information was shown in black, false information in red. One report contained false sentences that increased the severity of the crime, and the other included false sentences that diminished it. Some test participants were asked to do a concurrent digit-search task as they read the false sentences in the reports. This was meant to impose cognitive load. Afterwards, participants were asked what prison sentence (0-10 years) they would give for the crimes that they had read about on the first line and how they evaluate the criminal's character, for instance, how much they liked him, how dangerous he was, and how much counseling would help him.

It turned out that when the text contained exacerbating information that was false, subjects in the load condition, but not those in the no-load condition, recommended harsher sentences than when mitigating information was false. Furthermore, these participants' ratings of the perpetrator's dislikableness, dangerousness, and likelihood of benefitting from counseling were higher than those of the no-load subjects. Gilbert et al. and other Spinozans (Mandelbaum 2014; Levy and Mandelbaum 2014) argue that since the subjects under load acted on the false information just as if they believed it, they did indeed believe it. Since they seemed unable to suspend acceptance of the information, the findings suggest that subjects automatically believe the propositions they entertain, or so Spinozans claim.

In fact, they maintain that this will be the case even if subjects are *before* encountering the propositions told that the propositions are false. To support this, they cite a study by Wegener et al. (1985) in which participants were shown pairs of suicide notes and told that one note from each pair was real and the other fake (Gilbert 1991: 114; Mandelbaum 2014: 67; Levy and Mandelbaum 2014: 26). The subjects' task was to sort the real ones from the fakes. After each decision, they were given feedback on their performance. Crucially, before the trial started, they were informed that the feedback they would receive was false. After the test, subjects were asked to estimate how often they answered correctly.

Surprisingly, their answers still matched the feedback. Levy and Mandelbaum (2014) write that the

> knowledge of the feedback persists because the participants automatically affirm the feedback when they hear it, even though they know the feedback is false. Since they are engaged in a relatively fast-paced experiment, the participants lack the mental energy to override the false claims. (26)

On the basis of the data just introduced (and other findings), Spinozans claim that "when we hear someone speak [and think about what they are saying], normally what is said goes directly into belief" (Millikan 2004: 121), that "thinking is believing" (Mandelbaum 2014: 55), and that "we are designed to initially affirm any propositions that we happen to think about" (Levy and Mandelbaum 2014: 26). In the

next three sections, I'll motivate some doubts about these claims. I begin with a theoretical consideration.

3. What the data don't show

The just-mentioned studies, which play a pivotal role in the Spinozan argument, involve subjects that are under cognitive load or "lack mental energy" (Levy and Mandelbaum 2014: 26). The involvement of cognitive load or a depletion of mental energy is important for Spinozans because the empirical case for their theory rests on what Gilbert (1991: 109) calls a "general principle of systems break-down: *When stressed, modular information-passing systems with multiple exit capabilities will often show a bias toward prematurely outputting the products of early modules* [italics original].

Since cognitive load 'stresses' the modular information-passing systems involved in the comprehension of a proposition *p*, it should lead them to prematurely output the product of the module that processed *p* before the load occurred. Spinozans then predict that if one automatically accepted *p* when one is entertaining *p*, imposing cognitive load during the validation phase should induce one's cognitive system to prematurely issue the product of the earlier processor, i.e., an acceptance of *p*. The findings do support the prediction of an acceptance under load.

However, strictly speaking, they are compatible with the view that before cognitive load is imposed, the module processing *p* remains doxastically neutral about *p* and only at the moment when the load crosses a certain threshold opts for an acceptance of *p*. On this view, the acceptance of *p* that subjects exhibit in the mentioned studies is not the output of the module processing *p* before the validation, which is done by a different module. Rather, there is just one module responsible for both comprehension and validation and this module operates in stress conditions differently than in no-stress conditions: if subjects are during validation of *p* put under load, they will not remain doxastically neutral but accept *p*.

This does undermine the claim that we can always remain doxastically neutral when we are considering a proposition *p*, for sometimes we are considering *p* under load. But it does not show that when subjects are presented with *p* and not put under load, they will initially automatically believe the proposition. For all that the above studies tell us, when we are not under cognitive load or do not lack mental energy, we do not believe what we are thinking about until our mental resources are depleted. To retain the strong view that it is "(nomologically) impossible for one to not immediately believe propositions that one tokens", as Mandelbaum (2014: 61) claims, this possibility would need to be addressed and ruled out.[6]

[6] Gilbert (1991: 114f) considers the proposal that one might understand a proposition without representing it as true (114). In response, he cites Wegener et al.'s (1985) above-mentioned study in which subjects didn't refrain from accepting false propositions even though they were told about their falsity beforehand. Gilbert holds that "subjects were unable to represent the statements in a truth-neutral fashion, even when directly motivated to do so" (ibid: 115). However, this is unconvincing, as it might be that participants simply forgot to bring the relevant information on the falsity of the experimenter's statements (about their performance in identifying suicide notes) to bear on the issue, and took the experimenter to be a reliable source. Also, perhaps the test participants failed to resist acceptance because they are in the study "engaged in a relatively fast-paced experiment," and hence "lack the mental energy to override the false claims" (Levy and Mandelbaum 2014: 26). The findings then no longer undermine the proposal that when one's mental energy is *not* depleted, subjects can think about propositions without initially believing them. Gilbert (1991) offers another point in support of the claim that comprehension and acceptance always fall together. He reports a study in which he and his colleagues asked subjects to simply read out sentences on an imaginary creature without assessing the statements. Yet, when later on questioned about the veracity of the statements, subjects took them to be true. However, as Gilbert writes himself, subjects were asked to read quickly, and there was a premium on fast readers. Hence, test subjects were under time pressure. Since time pressure reduces mental energy, the findings again don't undermine the proposal that if subjects are not mentally taxed, they can think about a proposition without initially accepting it.

4. Automatic rejections

There is reason to hold that even when we *are* under cognitive load, we don't always initially automatically believe what we think about. According to the Spinozan theory, we automatically believe *any* proposition we entertain, yet, it is worth noting that unlike in, for instance, Gilbert at al.'s studies, in everyday life, people often have some knowledge available to draw on when they are confronted with a piece of information. Given this, suppose that we have strong background beliefs about a proposition *p* and these beliefs contradict *p*. Do we still initially automatically accept *p* when we entertain it?

Richter et al. (2009) conducted a version of Gilbert et al.'s experiment that did not use nonsensical statements such as 'A *monisha* is a star' but objectively true or false assertions about which subjects could be expected to have either strong or weak validity-related background beliefs. They found that for statements with strong background beliefs (true or false), say, 'Soft soap is edible', cognitive load during learning did not result in people's accepting false propositions. That is, when subjects were, after the learning phase, in the test phase asked 'Is soft soap edible?' then they didn't show evidence of an acceptance of the proposition.

Could it be that subjects simply accessed their stored strong background belief that soft soap is inedible to answer the question, and thus showed unimpaired accuracy even though the on-line effortful rejection process was disrupted and an initial automatic acceptance in the learning phase occurred? Richter et al. used two different measures to rule this out.

First, they included new assertions in the verification task in addition to those that subjects had been presented with in the learning phase. By comparing the error rate and response latency for new assertions and assertions presented in the learning phase, Richter et al. could delineate effects of validation processes in the learning phase and separate these effects from belief effects that occurred in the test phase.

Second, they asked subjects to make their verification judgments within a specified time frame that varied in length. The thought was that if background beliefs become operative during resource-dependent validation processes in the verification task, the verification of assertions linked with strong background beliefs should be negatively affected by a shorter response timeframe. This did in fact happen with *new* (strong background belief-related) assertions. But crucially, when subjects verified assertions that were linked to strong background beliefs *and* shown in the learning phase, the decline from the long to the short response-time frame was only moderate. This suggests that the validation of the assertions already occurred under load in the learning phase, and that subjects were able to automatically reject what they thought about, which is at odds with the Spinozan theory.

Richter et al. (2009) conducted a second study that also speaks against the theory. Participants were very briefly (300-600ms; see experiments 3 and 4) presented with three words (one-by-one), which formed an assertion that was either consistent or inconsistent with their background beliefs. In the critical trials, the participants' task was to quickly assess the correct spelling of the third word while it was presented to them.

Subjects committed fewer mistakes and needed less time to respond when words within true sentences (i.e., sentences that were in line with their background beliefs) were grammatically correct and when words within false sentences (i.e., sentences that were at odds with their background beliefs) were grammatically incorrect than in the two incongruent conditions (i.e., correct grammar/with false statements and incorrect grammar/with true statements). Subjects seemed to quickly validate and sometimes reject the sentences, and the outcome of their validation affected their spell checking.

Notice that they weren't allowed to answer whenever they wanted to but were prompted to respond quickly at a particular moment, which was the same moment at which the truth-value of the assertion was accessible to them (as the third word completed the assertion). Hence, at that moment, their mental energy for the valuation was depleted. If the rejection of assertions is, as Spinozans claim, resource-dependent, this should have disrupted subjects' rejection of them. But it didn't, as is evidenced by the fact that the validation outcome, which in some cases was a rejection, affected the latency and error rate of the spell check.

It might be proposed that since strong background beliefs were in place very little effort was required and invested for rejections. However, it is hard to see why subjects should have invested *any* effort in rejecting the assertions. For investing cognitive effort is generally something that a subject does deliberately in order to achieve some goal or other. Yet in the study subjects were not asked to nor had the goal to understand, let alone validate, the assertions. It is thus less plausible to assume that they nonetheless effortfully rejected some of them. It is more likely that they did so automatically, which contradicts the Spinozan theory.[7]

5. Doxastic neutrality

According to the Spinozan theory, there also shouldn't be cases where subjects remain doxastically neutral about a proposition.[8] But this claim too is arguably false.

For instance, Hasson et al. (2005) conducted an experiment in which they presented subjects with a person's face (e.g., of a smiling man) and a statement about the person shown (e.g., 'This person thinks that things turn out for the best'). They used three types of statements: true statements that were also indicated as true, false statements that were also indicated as false, and truth-unspecified statements that were not indicated as either true or false. Right after the presentation, participants were presented with a word (for 250ms) and had to quickly press a button to indicate whether it was an English word. On the critical trials, the word presented (e.g., 'optimist') was related to either the true or the false version of the sentence preceding it (e.g., 'This person thinks that things turn out for the best').

Hasson et al. reasoned that if subjects represent any statement they entertain as true then those who are shown truth-value unspecified statements should respond equally quickly to terms connected with the truth of the sentences (henceforth 'true-related words') in the lexical decision task following *both* true and truth-value unspecified statements. If subjects don't do so, then they should respond more slowly to true-related words following truth-value unspecified statements than following true statements.

Hasson et al. found that lexical decisions about true-related words were faster when the statement was indicated to be true than when its veracity was unknown or when it was false. So, for instance, the word 'optimist' was evaluated more quickly when the statement 'This person thinks that things turn out for the best' was marked as true of a person than when the statement was truth-value unspecified or marked as false,[9] suggesting that subjects don't always represent the statements that they entertain as true, but in some cases can remain doxastically neutral about them.

[7] The assumption that subjects did invest effort in rejecting assertions despite not having the goal to validate them is also at odds with the well-documented finding that the human mind is a "cognitive miser" in that it tries to avoid spending cognitive resources and tends to adopt mental short-cuts whenever it can (Fiske and Taylor 2013; De Neys et al. 2013).

[8] See Gilbert's (1991: 109), and Mandelbaum's (2014: 62) figures of the Spinozan models; there is no state of doxastic neutrality or suspended belief.

[9] There might be a priming effect of veracity-related terms on subsequent lexical decisions about statement-related word, but this isn't very plausible, as it is hard to see a semantic link between, for instance, 'true' and 'optimist'.

A different set of studies lends further support to this view. According to the Spinozan theory, as Gilbert (1991) puts it, "ideas whose truth" have been "ascertained through a rational", effortful "assessment procedure" are "represented in the mind in precisely the same way as" are ideas that have "simply been comprehended; only ideas" that are "judged to be false" are "given a special tag" (109). True information that one automatically accepts or, upon reflection, endorses remains "untagged" (ibid).

With this in mind, Nadarevic and Erdfelder (2013) conducted a source-memory study in which test subjects learned statements from three different sources, i.e., fictitious persons called 'Hans', 'Fritz', and 'Paul'. They were told that each of the three persons differed in credibility, which meant that their statements had different truth-values (Hans = 100% true; Fritz = 50% true and 50% false; and Paul = 100% false statements). Half of the test subjects were told about Hans', Fritz's, and Paul's credibility, and therewith of the truth-value of these people's statements, *before* they were presented with the statements (pre-cue group). The other half was informed about it afterwards (post-cue group).

On the basis of studies that show that source memory for validity information is superior to source memory for names (see Begg et al. 1992), Nadarevic and Erdfelder reasoned that pre-cue subjects should display better source memory than post-cue participants. Furthermore, if, as the Spinozan model predicts, people store only 'false' tags, then good source memory in the pre-cue condition should be limited to false statements.

Within the pre-cue group, source memory turned out to be equally good for the true and false statements and was much better than source memory for statements of uncertain validity. Unlike the Spinozan view predicts, subjects seemed to tag statements as true and could refrain from encoding statements as either true or false.[10] For if they had encoded the (uncertain) statements of the unreliable source automatically as true, then pre-cue subjects should have recalled the source of these statements as well as the sources of the true and false statements. But that is not what Nadarevic and Erdfelder found, which suggests that subjects can remain doxastically neutral about propositions.[11]

6. Does the automaticity of believing confer obligations?

So far I have tried to cast doubts on the Spinozan claim that we always initially automatically believe everything we think about. I now want to take a critical look at Levy and Mandelbaum's (2014) point that the automaticity of believing has implications for the ethics of belief. For the sake of argument, I shall set aside the counterevidence to the Spinozan theory that I've just reviewed.

On the basis of the empirical case for the Spinozan theory, Levy and Mandelbaum maintain that we "are designed to initially affirm any propositions that we happen to think about" (ibid: 26). They continue that, as a result, those of us "who know about" our "propensities" to believe propositions through merely

[10] It might be argued that in the study, subjects equally well recalled the sources of true and false statements because they had enough time to consciously endorse (and not merely to unconsciously automatically accept) statements from a reliable source, which is in line with the Spinozan view. However, this still doesn't explain why subjects were worse at recalling the source of statements with uncertain validity. For, on the Spinozan view, these statements too should have been encoded as true, just as the statements in Gilbert et al.'s studies were under load encoded, and later on recalled, as true.

[11] An interesting experiment by Street and Kingstone's (2016) provides further evidence for this view. They presented participants with short video clips of people that were either lying or telling the truth. After each clip, the word 'Truth' or 'Lie' was shown on the screen, indicating whether the person had told the truth or lied. In some cases, during the verification, participants had to press a button when they heard a tone ring out, which was meant to deplete their cognitive resources. Afterwards, subjects were again presented with the images of the person. Some subjects were asked whether s/he told the truth or lied (truth-lie forced choice condition). The other subjects could also respond that they were unsure as to whether s/he told the truth or lied. Street and Kingstone found that only subjects in the truth-lie forced choice condition automatically took the person to be telling the truth. Subjects who could respond by opting for 'unsure' didn't exhibit that tendency, which suggests that subjects are able to merely entertain information.

entertaining them have "obligations to take the risk of forming unjustified" and "immoral beliefs into account" when we expose ourselves to them (ibid: 28, 30).

Levy and Mandelbaum's thought is that we often have control over what ideas we encounter, for instance, we have control over what television channel we put on (for instance, *Fox News, BBC* etc.). And since we "make it likely that we will acquire beliefs by mere exposure to them", just "as we have obligations to take risks into account when we act, we have obligations to take the risk of forming unjustified and [...] immoral beliefs into account when we expose ourselves to them", Levy and Mandelbaum conclude (ibid: 30).

One crucial assumption underlying Levy and Mandelbaum's argument is that subjects who "know about [their] propensities to acquire doxastic states through merely entertaining propositions" will still have the tendency to automatically believe them (Levy and Mandelbaum 2014: 28). This assumption, however, isn't supported by the studies (nor arguments) that Spinozans, including Levy and Mandelbaum, have mentioned. All of the studies that Spinozans typically cite involve subjects that are unaware of the proposal that people automatically accept the propositions that they entertain. Thus, as it stands, Levy and Mandelbaum's argument contains a gap. It leaves open the intriguing possibility that a subject's self-awareness of the tendency to automatically accept the propositions that she entertains might disable that tendency.[12] Interestingly, similar interference effects are in related cases arguably not just possible but actual.

One relevant study comes from research on stereotype processing. Stereotype activation, just as Spinozan belief formation, is often taken to be unconscious and beyond the subject's control. To test this, Moskowitz and Li (2011) conducted an experiment in which they indirectly activated egalitarian goals in some of their subjects by asking them to write down a short description of a past failure at being egalitarian toward African American men. Moskowitz and Li's rationale was that

> [m]any models of goal selection [...] reveal that a goal is triggered when one contemplates failure in the goal domain; by a person detecting a discrepancy between their actual responses and a desired response. This discrepancy is said to produce a psychological tension that impels the organism to reduce the tension and approach the standard. (106)

Building on that thought, Moskowitz and Li then asked both the subjects that had just written on their past failures to be egalitarian and control participants to do a lexical-decision task, which is often used to test automatic stereotypes (Banaji and Hardin 1996). Following a brief presentation of faces of either Black or White men, which they were told to ignore, subjects had to decide as quickly as possible whether or not a string of letters formed an English-language word, which was either a stereotype-relevant term (e.g., 'crime', 'stupid', 'lazy' etc.) or control word (e.g., 'annoying', 'nervous', 'indifferent' etc.). Moskowitz and Li's idea was that if stereotypes are activated by the face-primes (e.g., a Black face), subjects thus primed should be faster to respond to stereotype-relevant words. And if stereotype control occurs, this effect should disappear and, due to inhibition, stereotype-relevant words should be reacted to more slowly after the presentation of faces of Black men.

Interestingly, unlike control participants, subjects with indirectly activated egalitarian goals *did* display stereotype control and inhibition in the lexical-decision task even though

[12] Being told that the propositions that one will be presented with are false is distinct from being told that one has the tendency to automatically believe what one is thinking about. Hence, even if the former isn't sufficient for subjects to suspend automatic acceptance (as some Spinozans might argue by using Wegener et al. 1985), the latter might still be sufficient. I motivate this view below.

> during targeted questioning in the debriefing, no participant expressed any conscious intent to inhibit stereotypes on the task, nor saw any of the tasks performed during the computerized portion of the experiment as related to the egalitarian goals they had undermined [by the reflection on past failures at being egalitarian] earlier in the session. The reaction time task was not consciously seen as a way to address an egalitarian goal or as having anything to do with stereotyping. (Moskowitz and Li 2011: 108)

Hence, subjects "can control stereotyping without knowing a stereotype or a goal exists. Consciousness is not required. One's wants, even implicit wants, can direct thoughts" (ibid).

Moskowitz and Li's findings are relevant to Spinozan belief formation and Levy and Mandelbaum's argument pertaining to the ethics of belief. For suppose a subject *S* comes to believe that she has the tendency to automatically accept everything she is told. It is fair to say that *S* will take this to be at odds with the way she *should* form beliefs; gullibility is usually criticised as epistemically problematic.[13]

Since that is so, there will for her be a discrepancy between her actual response to propositions and her desired one. As in the stereotype study, this discrepancy may then produce a psychological tension that impels her to reduce the discrepancy by trying to approach her normative standard.[14] If we use the results of Moskowitz and Li's stereotype study as a model, it is not unreasonable to suspect that *S* will then form the implicit goal to *not* automatically accept the propositions that she entertains, which might subsequently inhibit her tendency to form beliefs automatically, just as the implicit egalitarian goal in the stereotype study inhibited subjects' automatic stereotyping.

Of course, whether or not this is in fact the case remains to be seen. But my goal here is modest. It is only to add some plausibility to the view that an insight into one's automatic belief formation might interfere with the latter. Moskowitz and Li's stereotype study and the just mentioned line of thought do help motivate this, and therewith suggest that Levy and Mandelbaum's so far uncorroborated assumption that such an insight cannot have that effect is in need of further support.[15]

7. Conclusion

I argued that there is reason to doubt the Spinozan theory, according to which we always initially automatically believe what we think about. The cognitive load studies, which are one of the main sources of support for the theory, are compatible with the view that when we are not under cognitive load, we don't initially automatically believe the propositions that we entertain. There are also studies that suggest that sometimes we automatically reject propositions, or remain doxastically neutral about them.

Furthermore, I maintained that even if we set these studies aside and take the empirical case for the Spinozan theory at face value, Levy and Mandelbaum's (2014) argument that those of us who are aware

[13] For instance, Faulkner (2000) writes that given that a "speaker's intentions in communicating need not be informative and given the relevance of these intentions to the acquisition of testimonial knowledge", it is "doxastically irresponsible to accept testimony without some background belief in the testimony's credibility or truth" (587-8).

[14] This provides a response to the objection that if subjects *could* refrain from accepting the propositions that they entertain at all then surely when they are told before the presentation of some propositions that the latter will be false, they should refrain from accepting them (which, as Wegener et al.'s (1985) suggest, they don't do). The response to this point is that being told that the propositions will be false won't produce the psychological tension that is required for the mentioned interference with automatic processing.

[15] Levy and Mandelbaum might plausibly respond to the preceding point by holding that even people who have no idea that they form beliefs automatically are still responsible for their automatic belief formation. But this would require a different argument than the one that they currently propose.

of their automatic belief acquisition have new epistemic obligations is not fully convincing. For one of the assumptions that the argument rests on (i.e., the view that subjects' awareness of their tendency to form beliefs automatically leaves that tendency unaffected) is unsupported.

Nonetheless, Levy and Mandelbaum have rightly emphasised the importance of cognitive scientific findings on belief formation for ethical questions about how we should act when we expose ourselves to information. Because even if in subjects who believe that they tend to accept what they think about, this tendency is counteracted, the empirical findings on automatic belief formation do still confer one basic obligation onto us: to make sure that others – especially, for instance, judges and jury members in court, who ought to refrain from accepting (or rejecting) propositions unless the evidence supports doing so – know about the way they form beliefs. For this knowledge may play a critical role in enabling them to engage in impartial judgment- and decision-making.

References

Banaji, M., and Hardin, C. (1996). Automatic stereotyping. *Psychological Science*, 7(3), 136–141.

Begg, I., Anas, A., and Farinacci, S. (1992). Dissociation of processes in belief: Source recollection, statement familiarity, and the illusion of truth. *Journal of Experimental Psychology*, 121, 446–458.

Dennett, D. (1981). *Brainstorms: Philosophical Essays on Mind and Psychology*. Cambridge, MA: MIT Press.

De Neys, W., Rossi, S., and Houdé, O., (2013). Bats, balls, and substitution sensitivity: Cognitive misers are no happy fools. *Psychonomic Bulletin and Review*, 20, 269–273.

Descartes, R. (1641/1984). *Meditations on first philosophy: In which the existence of God and the distinction of the soul from the body are demonstrated*. Indianapolis: Hackett.

Faulkner, P. (2000). The Social Character of Testimonial Knowledge. *Journal of Philosophy*, 97, 581–601.

Fiske, S., and Taylor, S., 2013. *Social cognition*. London: Sage.

Fodor, J. (1983). *Modularity of Mind*. Cambridge, MA: MIT Press.

Gilbert, D., Krull, D., and Malone, P. (1990). Unbelieving the unbelievable: Some problems in the rejection of false information. *Journal of Personality and Social Psychology*, 59, 601–613.

Gilbert, D. (1991). How mental systems believe. *American Psychologist*, 46, 2, 107–119.

Gilbert, D., Tafarodi, R., and Malone, P. (1993). You can't not believe everything you read. *Journal of Personality and Social Psychology*, 65, 221–233.

Kriegel, U. (2013). Entertaining as a Propositional Attitude: A Non-Reductive Characterization. *American Philosophical Quarterly*, 50, 1–22.

Levy, N., and Mandelbaum, E. (2014). The powers that bind: Doxastic voluntarism and epistemic obligation. In: Matheson, J. (ed.), *The Ethics of Belief*. Oxford: Oxford University Press, 12–33.

Mandelbaum, E. (2014). Thinking is believing. *Inquiry: An Interdisciplinary Journal of Philosophy*, 57, 1, 55–96.

McDowell, J. (1998). Having the world in view: Sellars, Kant and Intentionality. Lecture 1: Sellars on perceptual experience. *Journal of Philosophy*, 95, 431–450.

McHugh, C. (2011). Judging as a Non-Voluntary Action. *Philosophical Studies*, 152, 245–69

Meissner, C., and Kassin, S. (2002). 'He's guilty!' Investigator bias in judgments of truth and deception. *Law and Human Behavior*, 26, 5, 469–480.

Millikan, R. (2004). *Varieties of Meaning.* Cambridge, Massachusetts: MIT Press.

Moskowitz, G. and Li, P. (2011). Egalitarian Goals Trigger Stereotype Inhibition: A Proactive Form of Stereotype Control. *Journal of Experimental Social Psychology*, 47(1), 103–116.

Nadarevic, L., and Erdfelder, E. (2013). Spinoza's error: Memory for truth and falsity. *Memory & Cognition*, 41, 176–186.

O'Brien, L. (2007). *Self-knowing agents*. Oxford: OUP.

Richter, T., Schroeder, S., and Wohrmann, B. (2009). You don't have to believe everything you read: Background knowledge permits fast and efficient validation of information. *Journal of Personality and Social Psychology*, 96, 538–558.

Spinoza, B. (1677/1982). *The Ethics and Selected Letters*. In S. Feldman and S. Shirely (Eds.). Indianapolis, IN: Hackett.

Street, C. and Kingstone, A. (2016). Aligning Spinoza with Descartes: An informed Cartesian account of the truth bias. *British Journal of Psychology*, 1–14.

Wegner, D., Coulton, G. and Wenzloff, R. (1985). The Transparency of Denial: Briefing in the Debriefing Paradigm. *Journal of Personality and Social Psychology* 49, 2, 338–46.