UNIVERSITAT DE BARCELONA

The representational structure of implícit attitudes

Ilia Patronnikov

UNIVERSITAT DE BARCELONA

Tesi doctoral

The representational structure of implicit attitudes

Autor/a:

Ilia Patronnikov

Director/s:

Josefa Toribio Mateas

Manuel Jesús Martínez Merino

Tutor/a:

Josefa Toribio Mateas

Programa de Doctorat

CIÈNCIA COGNITIVA I LLENGUATGE

Facultat de Filosofia

Maig de 2025

Contents

Resum - Català

Les actituds implícites (AI) són estats mentals que causen comportaments discriminatoris de baixa intensitat als quals ens solem referir com a "biaixos implícits". Per exemple, un cap de personal pot creure honestament que la raça no té cap relació amb la intel·ligència o la competència, però rebutjar sistemàticament les sol·licituds de persones d'una raça determinada. Aquesta conducta discriminatòria del cap de personal podria ser deguda a les seves AI vers les persones d'aquesta raça. Darrerament, les AI han estat un important focus d'interès per la psicologia i la filosofia. Això pot ser degut, en part, al fet que s'han desenvolupat procediments indirectes per mesurar aquestes actituds. Aquests nous procediments han permès als investigadors avaluar les AI dels individus cap a diferents grups socials, com ara grups racials o religiosos, sense haver de recórrer a qüestionaris d'autoavaluació.

En aquesta tesi s'abordarà una qüestió central en la investigació sobre les AI: Quina és la naturalesa de les AI? O, dit altrament, quin tipus d'estats mentals són les AI? Aquesta pregunta és interessant en si mateixa, però també té importants implicacions per a qüestions com ara si som moralment responsables de les nostres AI i què es pot fer per atenuar-les. En aquesta tesi es defensarà la teoria de les creences, una teoria d'acord amb la qual les AI són creences. L'estratègia consistirà a formular una objecció en contra d'aquesta teoria i mostrar com es pot respondre. A més, es tractarà de clarificar una noció clau en els arguments a favor de la teoria, la noció d'inferència, reforçant així els fonaments conceptuals del debat sobre les AI.

La tesi s'estructura de la següent manera. En el capítol 1, es descriu el fenomen de les AI i els desafiaments teòrics que planteja. Després de presentar els mètodes empírics utilitzats per estudiar les AI, s'analitzaran les diferents propostes teòriques que tracten de donar compte de la seva naturalesa i s'examinarà l'evidència empírica que apunta al fet que les AI són creences. En el capítol 2 es discutirà el que anomenarem "objecció de l'autoconeixement" a la teoria de les creences: les persones solen tenir una bona posició epistèmica pel que fa a les seves

creences; no obstant això, sovint, desconeixen les seves AI. Però si les AI són creences, com s'explica aquesta asimetria? S'argumentarà que les AI constitueixen un tipus especial de creences, creences que no van acompanyades dels judicis corresponents, i que aquests judicis són una via important per conèixer les creences d'un mateix. Atès que aquesta via està bloquejada per les AI, això pot donar compte de l'asimetria epistèmica. El capítol 3, se centra en la noció d'inferència. Un argument important a favor de la teoria de les creences sosté que les AI apareixen en les transicions inferencials, cosa que indica que són creences. Per avaluar aquest argument, cal aclarir què és una inferència. Es defensarà que hi ha dos enfocaments per comprendre les inferències. D'acord un primer enfocament, al qual ens referirem com "enfocament dels processos mentals", una inferència es defineix en termes de l'estructura psicològica subjacent a les transicions entre estats mentals. Per contra, el segon enfocament, – l'anomenarem "epistèmic" – explica la noció de transició inferencial en termes epistèmics, sense fer referència a l'estructura dels processos mentals que la fonamenten. Ens centrarem en una explicació epistèmica de la inferència – la Hipòtesi de Resposta (Response Hypothesis) de Susanna Siegel – i s'argumentarà que aquesta permet que estats amb estructura associativa apareguin en les transicions inferencials. Amb aquesta discussió es pretén mostrar que, perquè l'argument basat en la inferència a favor de la teoria de les creences sigui vàlid, cal adoptar el primer dels enfocaments, l'enfocament dels processos mentals, per caracteritzar la inferència.

Paraules clau: actituds implícites, biaix implícit, creença, autoconeixement, inferència

Resumen – Español

Las actitudes implícitas (AI) son estados mentales que causan comportamientos discriminatorios sutiles, a menudo denominados "sesgos implícitos". Por ejemplo, un responsable de contratación puede creer sinceramente que la raza es irrelevante para la inteligencia y la competencia, pero rechazar sistemáticamente las solicitudes de miembros de una determinada raza. Dicha discriminación podría deberse a la AI del responsable hacia dicha raza. En los últimos años, las AI han atraído una atención considerable en psicología y filosofía, en parte debido al desarrollo de medidas indirectas de actitudes, que permiten a los investigadores evaluar la actitud de los individuos hacia grupos sociales, como grupos raciales o religiosos, sin depender de cuestionarios de autoinforme.

Esta tesis aborda la siguiente pregunta central en la investigación sobre las AI: ¿Cuál es la naturaleza de las AI? O, dicho de otro modo, ¿qué tipo de estados mentales son las AI? Esta pregunta es interesante en sí misma y tiene importantes implicaciones para cuestiones como la responsabilidad moral asociada a las AI y su mitigación. En resumen, defiendo la teoría según la cual las AI son creencias; llamémoslo la perspectiva de las creencias. Más precisamente, formulo una objeción a esta perspectiva y muestro cómo puede responderse. Asimismo, aclaro una noción clave en los argumentos a favor de esta teoría: la noción de inferencia, reforzando así los fundamentos conceptuales del debate sobre las AI.

La tesis se desarrolla de la siguiente manera. En el capítulo 1, describo el fenómeno de las AI y los desafíos teóricos que plantean. Tras introducir los métodos empíricos utilizados para estudiarlas, analizo las teorías contrapuestas sobre su naturaleza y examino la evidencia empírica que sugiere que las AI son creencias. El capítulo 2 aborda lo que llamo “la objeción del autoconocimiento” a la perspectiva de las creencias: las personas suelen tener una buena posición epistémica con respecto a sus creencias; sin embargo, a menudo, desconocen sus AI. Pero si las AI son creencias, ¿cómo se explica esta asimetría? Argumento que las AI constituyen

un tipo especial de creencia: creencias que no van acompañadas de los juicios correspondientes, y que tales juicios son una vía importante para conocer las propias creencias. Dado que esta vía está bloqueada para las AI, se puede explicar la asimetría epistémica. El capítulo 3 se centra en la noción de inferencia. Un argumento importante a favor de la perspectiva de las creencias sostiene que las AI aparecen en las transiciones inferenciales, lo cual indica que son creencias. Para evaluar este argumento, es necesario aclarar qué se entiende por inferencia. Sostengo que existen dos enfoques para comprenderla. En uno, que denominaré "enfoque de los procesos mentales", la inferencia se define en términos de la estructura psicológica que subyace a las transiciones entre estados mentales. Por el contrario, el otro enfoque —llamémoslo "epistémico"— explica la noción de transición inferencial en términos epistémicos, abstrayéndose de la estructura de los procesos mentales que la sustentan. Me centro en una explicación epistémica de la inferencia —la Hipótesis de Respuesta (Response Hypothesis) de Susanna Siegel— y argumento que permite la posibilidad de que estados con estructura asociativa aparezcan en transiciones inferenciales. Esta discusión muestra que, para que el argumento basado en la inferencia a favor de la perspectiva de las creencias sea válido, es necesario adoptar el enfoque de los procesos mentales para caracterizar la inferencia.

Palabras clave: actitudes implícitas, sesgos implícitos, creencia, autoconocimiento, inferencia

Abstract - English

Implicit attitudes (IAs henceforth) are mental states that cause subtle discriminatory behaviors often referred to as "implicit bias". For example, a hiring manager may sincerely think that race is irrelevant to intelligence and competence, yet systematically reject applications from members of a certain race. Such discrimination might be caused by the manager's IA towards that race. In recent years, IAs have attracted significant attention in psychology and philosophy, partially due to the development of indirect measures of attitudes, which allow researchers to assess individuals' attitude towards social groups, such as racial or religious groups, without relying on self-report questionnaires.

This thesis addresses the following central question in IA research: What is the nature of IAs? Or, differently put, what kind of mental states are IAs? This question is interesting in its own right and has important implications for issues such as moral responsibility for IAs and their mitigation. In a nutshell, I defend the belief view – the view that IAs are beliefs. More precisely, I formulate an objection to this account of IAs and show how it can be answered. I also clarify a notion that plays a key role in arguments for the belief view – the notion of inference – thereby strengthening the conceptual foundations of the debate over IAs.

The thesis proceeds as follows. In Chapter 1, I outline the phenomenon of IAs and the theoretical challenges they raise. After introducing the empirical methods used to study IAs, I discuss competing accounts of their nature and examine empirical evidence suggesting that IAs are beliefs. Chapter 2 addresses what I call "the self-knowledge objection" to the belief view: People are typically in a good epistemic position with respect to their beliefs, but they are often unaware of their IAs. But if IAs are beliefs, how can this asymmetry be explained? I argue that IAs constitute a special kind of belief – beliefs not accompanied by the relevant judgments and that judging is an important route to knowing one's beliefs. Since this route is blocked for IAs, the epistemic asymmetry can be explained. Chapter 3 focuses on the notion of inference. An

important argument for the belief view holds that IAs feature in inferential transitions, thus indicating that they are beliefs. To assess this argument, we must clarify what inference is. I argue that there are two approaches to understanding inference. On one view, which I will call *"*the mental-processes approach*"*, inference is defined in terms of the psychological structure underwriting transitions between mental states. By contrast, the other approach – let's call it "epistemic" – spells out the notion of inferential transition in epistemic terms while abstracting away from the structure of the mental processes underwriting them. I focus on one epistemic account of inference – Susanna Siegel's Response Hypothesis – and argue that it allows for the possibility that states with associative structure feature in inferential transitions. This discussion shows that, if the inference-based argument for the belief view is to go through, one must adopt the mental-processes approach to characterizing inference.

Keywords: implicit attitudes, implicit bias, belief, self-knowledge, inference

Acknowledgments

Writing a thesis is a challenging task. I suppose almost every PhD student would agree with this claim. Here is a description of the average PhD journey: Very often, one starts with vaguely formulated research objectives. To sharpen them, one reads tons of joyless papers, many of which turn out to be useless. Once the objectives are clarified, one starts writing one's own paper with the hope of making a teeny-tiny contribution to the field. Unfortunately, there is often no way of knowing whether one has succeeded until one sends the paper to a journal and gets some feedback. However, waiting for feedback may take forever, and in some cases, the critique isn't really helpful. Understandably, many people get frustrated and start wondering whether their thesis is worth finishing.

I was lucky to be part of the research environment that helped me to mitigate many of these problems (perhaps all of them, except for the unduly long review process). I am grateful to numerous people who've been with me throughout the PhD years and made writing the thesis not only a feasible but also fun enterprise.

I would like to thank the LOGOS Research Group in general. It is an amazing place for doing analytic philosophy.

I am grateful to the following people with whom I had a few exciting reading groups in the philosophy of mind: Filippo Contesi, Esa Díaz-León, Daniel Gregory, Marta Jorba, Joshua Shepherd, Thomas Sturm and Víctor Verdejo.

I would also like to thank the epistemology team of LOGOS: Fernando Broncano-Berrocal, Dario Mortini, Michele Palmira and, particularly, Sven Rosenkranz. With their help, I've come to know things that I wasn't in a position to know (nor justifiably believe) before.

I also wish to thank my fellow PhD students: Niccolò D'Agruma, Daphne Bernués, Marcelino Botín, Duccio Calosi, Giovanni Dusi, Lucía González Arias, Margherita Grassi, Markel

Kortabarria, Marc Lara Crosas, Marc Mela Quílez, Simone Melis, Andrea Rivadulla Duró, Niccolò Rosi and Adrián Solís Peña (sorry if forgot to mention someone!).

I've greatly benefited from a research stay at the Centre for Philosophical Psychology in Antwerp. Thanks so much to Bence Nanay and his crew: Adriana Alcaraz, Julian Bacharach, Stephen Gadsby, Ben Henke, Magdalini Koukou, Kyle Landrum, Kael McCormack-Skewes, Rebecca Rowson and Francesca Secco.

But most of all, I am indebted to my supervisor Pepa Toribio. It is difficult to overestimate her influence on my formation as a philosopher.

Finally, I wish to thank by family – my parents Svetlana and Andrey, my wife Julia and the little Mark and Gabriel. Their emotional support has been invaluable.

Research for this work was supported by pre-doctoral contract PRE2019-089063 funded by MICIU/AEI/ 10.13039/501100011033 and by ESF Investing in your future, by grants PGC2018-095909-B-100, Awareness, self-awareness and unawareness: Exploring the perception-cognition-action continuum (PI: Josefa Toribio) and PID2021-124100NB-I00, Philosophy of Social Cognition (PIs: Josefa Toribio and Esa Díaz León), funded by MICIU/AEI/ 10.13039/501100011033 and by ERDF A way of making Europe, by grant CEX2021-001169-M, María de Maeztu Units of Excellence funded by MICIU/AEI/10.13039/501100011033, and by AGAUR, under grant agreement 2021-SGR-00276 (LOGOS. Research Group in Analytic Philosophy).

Chapter 1: Introduction

Consider the following situation: A hiring manager is reviewing the CVs of two job candidates, A and B. Both applicants look equally strong and suitable for the position. The manager struggles to decide between them. Suddenly, certain details in A's CV catch his attention and make him think that this candidate might belong to race R. However, the manager quickly dismisses this thought and proceeds with the hiring process. After some deliberation, he chooses candidate B. When asked to justify his decision, he says that B has a bit more relevant work experience than A. It never occurs to him that A's perceived race might have influenced the decision: after all, the manager sincerely thinks that race is irrelevant to intelligence and competence. However, his hiring record reveals a disturbing pattern: he consistently rejects candidates of race R, even when they are just as qualified for the job.

It is not easy to describe the manager's attitude towards people of race R. On the face of it, he doesn't believe that they are less intelligent or less competent. When asked his opinion about these matters, he sincerely says that race doesn't affect competence or intelligence. Nonetheless, his hiring decisions are not in line with this belief. It seems that, at some level, he does think that people of race R *are* less intelligent and less competent, and this thought influences his decision making. The term "thought" may be vague, but it captures an essential feature of the situation: the manager has a certain mental state, and this mental state is responsible for his discriminatory hiring decisions.

To better understand such cases, we need a more precise characterization of mental states responsible for this kind of discriminatory behavior. They are often referred to "implicit attitudes" (IAs henceforth). There are many interesting philosophical questions that one can ask about IAs: Are individuals morally responsible for actions influenced by IAs? Do they have access to IAs, and if so, what kind of access? What interventions can help to reduce or eliminate

IAs? This focuses on a different question: What is the nature of IAs? Or, differently put, what kind of mental states are IAs?

My contribution to this debate is the following. First, I articulate a new challenge to one view about the nature of IAs – the view holding that IAs are beliefs. Second, I show how proponents of the belief view can respond to the challenge, and thereby lend it additional support. Finally, I clarify a central notion in arguments supporting the belief view – the notion of inference. I isolate two approaches to characterizing inference and argue that, if these arguments are to go through, they must rely on only one of these approaches.

Here is the plan for the remainder of this Introduction. Section 1 presents the research questions that this thesis addresses. Section 2 outlines the research objectives. Section 3 discusses the methodology employed throughout the thesis. In Section 4, I explain how IAs are measured and clarify some crucial terms. In Section 5, I present competing views about the nature of IAs. In Section 6, I discuss evidence for the belief view and clarify the notion of belief that is at work in this debate. Section 7 introduces my challenge to the belief view and outlines my response to it. Section 8 examines the competing approaches to characterizing inference and explains their relevance to the debate about IAs. These last two sections are based on peer-reviewed articles, which constitute Chapter 2 and Chapter 3 of this thesis. The concluding chapter presents the research findings, examines some remaining open questions and provides a summary of the thesis.

Section 1: Research questions

This thesis focuses on the issue of the nature of IAs. Thus, the main research question addressed in the research can be formulated as follows:

What kind of mental states are implicit attitudes?

I defend the view that IAs are beliefs. That is, I argue in favor of the belief view of IAs. My strategy is twofold. First, I examine empirical data suggesting that IAs tend to behave like ordinary beliefs. For instance, some studies indicate that IAs update in response to many of the same factors that drive belief revision. Second, I compare the belief view with competing accounts, arguing that it best explains the available empirical evidence. This strategy prompts a further question:

What notion of belief is at work in the debate about the nature of IAs?

This notion, which often remains implicit, is a combination of two ideas. First, most proponents of the belief view typically subscribe to a representationalist framework: they hold that believing requires having a mental representation with propositional content. Consider, for instance, the belief that Paris is the capital of France. It involves a mental representation, whose content is the proposition *Paris is the capital of France*, usually expressed in a declarative sentence. However, merely having a representation with propositional content is not sufficient for belief, since one might have different propositional attitudes (suspecting, desiring, etc.) to the same content. To distinguish belief from other propositional attitudes, one usually appeals to a second idea: functionalism. According to functionalism, belief is defined by the cognitive role it plays in one's cognitive economy – by how it interacts with other mental states and behavior. This functionalist lens aligns with how researchers adjudicate between competing accounts of IAs: they look at how IAs behave and, based on that, make conclusions about the type of mental state they belong to.

Evidence supporting the belief view also raises the following question:

How can we explain the differences between IAs and ordinary beliefs?

A common objection against the belief view follows this pattern: IAs have property F, ordinary beliefs lack F, therefore, IAs are not beliefs. A salient case concerns self-knowledge: Typically,

if a person believes that P, they are in a position to know (or at least justifiably believe) that they believe P. But empirical research shows that people are often unaware of their IAs. If IAs are beliefs, how could this asymmetry be explained? This leads to the following question:

What accounts for the asymmetry between implicit attitudes and ordinary belief with respect to self-knowledge?

Finally, answering the main research question about the nature of IAs leads me to clarify a notion that plays a key role in arguments for the belief view – the notion of inference. An important argument for the belief view holds that IAs figure in inferential transitions, thus indicating that they are beliefs. To assess this, we must ask:

What is inference?

I argue that there are two major approaches to characterizing inference and that only one of them supports the belief view. Clarifying this distinction helps to avoid unproductive terminological debates and strengthens the conceptual foundations of the debate over IAs.

Section 2: Objectives

The thesis is structured around three core objectives.

First, it aims to provide a philosophical account of the nature of IAs. This ontological inquiry has implications for other philosophical issues. The ontology of IAs is tightly linked to moral responsibility over behaviors caused by these mental states. An influential view of responsibility ties it to control, and the degree to which the agent can control her IAs clearly depends on what kind of mental state they are.

The nature of IAs also bears on epistemic questions: Is the agent in a position to know her IAs? If so, what mechanisms underlie this access? Competing accounts of IAs have different

consequences for the accessibility issue: if mental states of a certain kind are inaccessible to the agent and IAs belong to this kind, one should conclude that IAs are inaccessible. This interplay between metaphysical and epistemological questions is reflected in my discussion of self-knowledge of IAs.

Second, the thesis aims to clarify the dialectical landscape of the IA debate. The goal is to examine different kinds of arguments in support of the belief view. I distinguish two common argumentative strategies. Strategy 1: Proponents of the belief view present empirical evidence and argue that only this view can explain it. Strategy 2: Opponents highlight a feature that distinguishes IAs from beliefs; defenders then explain how IAs can still be beliefs despite this difference.

I take a close look at both kinds of arguments and assess the state of the debate through the lenses of this distinction. I suggest that the second kind of arguments should be given more attention for the following reason: Current empirical evidence clearly indicates that IAs have many belief-like features, so a new bit of evidence demonstrating that IAs behave like beliefs might not significantly affect the overall dialectic. By contrast, explaining the few but important ways in which IAs differ from beliefs could decisively support the belief view.

Third, the thesis aims to strengthen the conceptual foundations of the debate over IAs. I do so by focusing on the notion of inference, which plays a prominent role in arguments supporting the belief view. The fact that IAs may feature in inferential transitions is often cited as evidence supporting the belief view. However, the notion of inference at work here is rarely clarified. The thesis aims to rectify this omission by discussing different approaches to inference and clarifying their relevance to the IA debate.

Section 3: Methodology

This research employs standard philosophical methods: conceptual analysis of the key notions featuring in the debate, critical assessment of arguments formulated by proponents of different accounts and rational reconstruction of their reasoning. IAs are a complex phenomenon that calls for each of these methods. To begin with, the very term "attitude" is open to several interpretations: it can refer to behavioral and cognitive dispositions pertaining to a given social group, and it can also denote mental states responsible for such dispositions. Unless these meanings are clearly distinguished, there is room for misunderstanding and error.

The same applies to other key notions – belief, association, inference. Consider the notion of association. On one reading of the term "association", all mental states are associations, since all cognition may be implemented by connectionist networks made up of nodes connected via links. On another reading, this term refers to a specific kind of non-propositional structure, in which case some mental states – states with propositional content – are not associations. Recognizing this distinction helps to avoid equivocation.

Adjudicating between different accounts of IAs requires weighing their advantages and disadvantages. This, in turn, involves a clear understanding of arguments on which the accounts are based. Some arguments are inferences to the best explanation, others are of a more conceptual nature. Each type of arguments needs to be addressed in its own way. For example, the claim that IAs are beliefs because they are sensitive to such epistemic factors as the perceived reliability of the source of information or the believability of a message is an instance of inference to the best explanation. This argument works only if it can be shown that competing accounts of IAs can't accommodate responsiveness to epistemic factors. By contrast, the claim that IAs are inaccessible to the agent because they are unconscious beliefs

is a more conceptual argument. It requires elaboration of the notion of unconscious and determining whether this notion is applicable to the phenomenon in question.

Finally, the methodology is not confined to philosophical methods – it also incorporates empirical research. IAs have been extensively studied in psychology and cognitive science through various experimental paradigms. This experimental work is especially relevant for the question of the nature of IAs: it is a common assumption that different types of mental states are defined by their functional profiles. So, determining the type to which IAs belong requires examining how they behave. For this reason, empirical literature can't be ignored. A significant part of the thesis engages with this literature, interpreting its implications. I describe different studies and explain their relevance for the central question of the thesis.

Section 4: How to measure implicit attitudes?

The hiring scenario described above seems plausible. Many of us would agree that similar cases happen in real life. However, imaginary scenarios only get us so far: they leave many features of IAs unspecified, and without a clear understanding of these features, we can't determine what IAs are. Are they unconscious beliefs? Are they associations between social categories and negative attributes? Answering these questions requires empirical data, not just thought experiments. We must examine how IAs behave in real life. And to do so, we must know how to measure them.

How do we go about investigating people's attitude towards, say, a particular race? Traditional measures of attitudes in psychology rely on self-report: participants are given questionnaires, and their answers are taken to reveal their attitudes. But direct, self-report-based measures are clearly inadequate in cases of subtle discrimination like the hiring scenario. In such cases, the

agent reports a positive attitude (or at least the absence of a negative one) but behaves in a way that suggests otherwise.

In recent years, psychologists have developed a number of indirect tests designed to capture such mental states that are difficult to report. These include the Implicit Association Test (Greenwald et al. 1998), evaluative priming (Fazio et al. 1986), Affect Misattribution Procedure (Payne et al. 2005), Go/No-Go Association test (Nosek and Banaji 2001), to name a few. Here is a simplified illustration of how the Affect Misattribution Procedure (AMP) works: A participant sees an affect-laden image (a kitten, say). This image is followed by a neutral target object (a Chinese character, say). The task is to classify the target object as pleasant or unpleasant while ignoring the influence of the affect-laden prime. It turns out that, despite these instructions, the prime influences the evaluation of the target object. For example, a positive prime makes one categorize a neutral object as pleasant. The influence of the prime on the target evaluation can be used as an index of the attitude towards the prime.

This kind of data is trivial when the prime is a kitten – we don't need the AMP to determine that people like kittens. But imagine that we want to know one's attitude towards a stigmatized social group. A participant's answers to a questionnaire indicate a lack of a negative attitude, but we have reasons to doubt this result. In such situations, the AMP becomes a valuable tool by giving us access to an otherwise inaccessible attitude. We can use a picture of a person of a stigmatized social group as the prime and see whether the picture affects the evaluation of neutral target objects. Suppose the evaluation is shifted in the negative direction. This indicates the participant might have a negative IA to the race in question. The magnitude of the shift serves as an index of the strength of this attitude.

Other indirect tests work in a similar way: participants are given a task, and some features of their performance (the evaluation of the target object, the reaction time, the number of errors,

etc.) are interpreted as indicators of an underlying attitude. These tests have been widely used to measure attitudes to races, religious groups, genders and sexual orientations, to name a few domains of application. Their indirect nature is useful for two reasons. First, race, religion, etc. are sensitive topics for many people, so they might not be completely frank when filling out a questionnaire. Second, as the hiring example shows, people might be mistaken about their attitudes. The indirect measures help to overcome these two difficulties[1].

Let me end this section with terminological clarification. IAs are typically contrasted with explicit attitudes (EAs henceforth). What's the difference between these two kinds of attitudes? As I will use these terms, "IAs" refers to mental states measured by indirect methods like the AMP, whereas "EAs" denotes mental states tapped into by direct measures like questionnaires (see Greenwald 2020, Kurdi 2019 for this usage). This definition doesn't say anything substantive about IAs, which is an advantage. The features of IAs should be determined empirically, not assumed at the outset. For this reason, they shouldn't be built into the definition[2].

That said, one shouldn't rule out that direct and indirect measures sometimes tap into the same mental state. If we give an open racist a questionnaire about his racial attitudes, the questionnaire might show that he dislikes people of a certain race (assuming that the person is being frank). An indirect test might also reveal a negative attitude to this race. In such cases, both tools might be measuring the same mental state. If so, the implicit and explicit attitudes of the person are constituted by the same mental state.

[1] The validity of indirect tests is debated. Some meta-analyses indicate that such measures poorly predict real-world behavior (Oswald et al. 2013). However, other meta-analyses suggest otherwise (Greenwald et al. 2015, Kurdi et al. 20018).

[2] Contrast this definition with one early characterization of IAs: IAs are object-valence associations in memory (Fazio 1995). This is inadequate because, as we will see, it is hotly debated whether IAs are associations. Or consider another definition: IAs are introspectively unidentified traces of past experience that mediate some responses (Banaji and Greenwald 1995). Again, it is an open question whether IAs are open to introspection.

But such alignment is rare. One robust empirical finding is that IAs and EAs are weakly correlated, though the correlation tends to be positive (Greenwald et al. 2020). It's hard to explain this finding assuming that direct and indirect tests track the same mental states. A natural way to accommodate it is to say that people often have conflicting IAs and EAs. Thus, scenarios similar to the hiring situation seem to be widespread.

One final note: the term "implicit bias" is sometimes used to refer to mental states responsible for subtle discrimination, and sometimes, to the behaviors triggered by IAs. In this thesis, I use it in the latter sense: to denote discriminatory behavior caused by IAs.

Section 5: Different views about the nature of implicit attitudes

We can now return to the central question of the thesis: What are IAs? There are many answers to offer. The goal of this section is to outline different accounts of the nature of IAs. I will also explain why the question is worth asking.

Up to this point, I've assumed that IAs are mental states, but not everybody agrees with this claim. For example, Edouard Machery (2016) argues that IAs are akin to character traits, such as courage. Being courageous amounts to having certain behavioral and cognitive dispositions and abilities: for instance, the ability to keep your fear under control in dangerous situations, the disposition to stand up to threats, and so on. Clearly, these dispositions depend on mental states: keeping the fear under control, arguably, requires cognitive control, and cognitive control involves, among other things, an internal representation of current task demands (Dixon 2015). But courage is not identical to any of these mental states because the same disposition can be realized by different mental states. On Machery's account, indirect tests of attitudes track various evaluative dispositions towards target stimuli, and it is a mistake to look for a common underlying mental state (or a cluster of mental states) causing the dispositions. Such

mental states may vary both across and within individuals. A similar dispositionalist approach is found in Eric Schwitzgebel's (2002) account of belief, which can be extended to IAs.

Another broadly dispositional account is Gabbrielle Johnson's (2020) functional view, which characterizes IAs as mental structures underwriting social-kind inductions. These structures are functionally described: they take certain mental states ("John belongs to race R") as inputs and yield as outputs some other mental states ("John is incompetent"). Crucially, the account leaves open which states and processes bridge the gap between inputs and outputs: it focuses on the functional role rather than on the nature and content of the intermediate states.

Other accounts are more specific about the nature of IAs: they identify such attitudes with a certain kind of mental state. For a long time, the debate was dominated by associationist accounts (Fazio 1995, Gawronski and Bodenhausen 2006, Holroyd 2016, Toribio 2018). This is reflected in the very name of one of the most popular indirect tests: the Implicit *Association* Test (IAT). Associationist accounts vary in detail[3], but they all subscribe to the claim that IAs are constituted by mental states with associative structure. A toy example of such a state (let's call it "an association" for simplicity) is a compound of the representations of salt and pepper[4]. Salt and pepper often go together (more technically, they are spatiotemporally contiguous). As a result, the ideas (or concepts) of salt and pepper get linked in such a way that the activation of one leads to the activation of the other. By the same token, a representation of a social group

[3] For example, the systems of evaluation (SEM) model posits two independent systems of mental representation, associative and rule-base (McConnell and Rydell 2014). By contrast, the Associative-Propositional Evaluation Model (APE, Gawronski and Bodenhausen 2006) and Meta-Cognitive Model (MCM, Petty et al. 2007) posit different kinds of *processes* drawing on a common associative representational store.

[4] The term "association" and its derivatives are used in several ways (Mandelbaum 2022). It can refer to a certain kind of mental structure – roughly, concepts and valences connected via excitatory and inhibitory links. It also denotes a type of learning – the kind of learning studied by Pavlov, who discovered that paring food with a neutral stimulus (say, a bell) makes it so that the neutral stimulus alone starts provoking the same response (salivation) as food. Finally, the term can refer to the structures implementing cognition. I use the word "association" in the first sense – to talk about a certain kind of mental structure.

may become associated with particular attributes or valences, giving rise to complex networks where the activation of one node spreads to the other.

The main competitor of associationism is the belief view, which holds that IAs are mental state of the same kind as ordinary beliefs, like the belief that Paris is the capital of France (De Houwer 2014, Mandelbaum 2016, Carruthers 2017). I will say more about the notion of belief in the next section.

Other views hold that IAs are imaginative or quasi-perceptual in nature. Bence Nanay (2021) argues that IAs are constituted by mental imagery: perceptual representations that are not directly triggered by sensory input. In a similar vein, Ema Sullivan-Bissett (2019) defends the view according to which IAs are unconscious imaginings. By "unconscious" she means states not available to introspection. Imaginings are mental states that, unlike beliefs and perceptual states, are not connected to truth. Some imaginings have propositional contents with makes them more similar to beliefs. Other imaginings might be imagistic and thus are more akin to perceptual states. Finally, there might be propositional imaginings. This category of imaginings has the same properties as mental imagery from Nanay's account.

Some accounts treat IAs as hybrid states. Tamar Gendler (2008) introduces the notion of aliefs – states composed of representational, affective and behavioral components. On this account, seeing a member of a certain racial group activates the representational component of an alief, and this in turn gives rise to certain feelings (say, unease) and behaviors (say, avoidance).

Michael Brownstein (2018) also holds that IAs are made up of different components. These components are abbreviated as "FTBA": perception of a *F*eature in the environment might result in bodily *T*ension. The tension triggers a *B*ehavioral response that is supposed to *A*lleviate the tension. This is very similar to the alief account: both views stress that IAs are made up of several parts – representations, feelings, behaviors – and that these parts are tightly linked.

Neil Levy (2015) argues that IAs are a sui generis kind of mental states, which he calls "patchy endorsements". Patchy endorsements are similar to beliefs in some respects and different from them in others. Like beliefs, IAs have propositional contents. This allows them to feature in relatively sophisticated mental transitions. However, Levy thinks that IAs are not sensitive enough to evidence and to the contents of other states to qualify as beliefs.

To complicate matters further, Jules Holroyd and Joseph Sweetman (2016) argue that implicit attitudes (implicit bias in their terminology) are heterogeneous states. They identify two dimensions of heterogeneity. First, they argue that IAs are constituted by associations with different types of content – roughly, semantic and affective. These two kinds of associations have different functional profiles, so the generalizations that apply to one kind may not be applicable to the other. Second, IAs might feature in different cognitive processes. For this reason, the behavior of the same attitude varies depending on what processes are recruited. One can add one more dimension of the heterogeneity and argue that IAs are constituted by different *types* of mental states. For example, some IAs may be associations and others, beliefs.

The question about the nature of IAs is not only interesting in itself – it bears on other important issues. For example, it is relevant for the issue of moral responsibility for actions influenced by IAs. Arguably, responsibility is tightly linked to control: more precisely, the agent is morally responsible for actions that are under her control (Levy 2014). Competing views about the nature of IAs have different implications for the controllability of such attitudes. For instance, the agent seems to have more control over her beliefs than over her associations. Associations are sensitive to spatiotemporal contiguity of stimuli, and people often have little influence over co-occurrences in their environment. If one is surrounded by annoying people who speak with, say, an Eastern-European accent, one may end up associating people from Eastern Europe with something negative. In these circumstances, there is very little that one can do to prevent the formation of this association. By contrast, one has some control over one's beliefs. For instance,

one can decide what media to follow, and this decision indirectly influences the content of one's beliefs.

Determining the nature of IAs is also important for IA mitigation. IAs cause subtle discriminatory behaviors, so combating discrimination would require the eradication, or at least weakening, of such attitudes. Different views about IAs provide different recipes for doing so. Again, we can contrast associationism and the belief view as an illustration. If IAs are associations, the most effective way to change a negative IA to a social group is to repeatedly pair stimuli connected to this group with something positive. By contrast, if IAs turn out to be beliefs, this strategy might not work: after all, belief revision is responsive to evidence, and the fact that a social group co-occurs with positive stimuli doesn't necessarily support the claim that its members have positive traits.

Section 6: The notion of belief and evidence for the belief view of implicit attitudes

Despite the complexity of the landscape described above, two views seem to dominate the debate – associationism and the belief view. In this section, I examine some of the evidence supporting the belief view. To do so, I begin by clarifying the notion of belief that is at work in the debate. I then briefly explain why the evidence poses challenges for the competing accounts.

What does it mean to believe that P, where P stands for a declarative sentence like "Paris is the capital of France"? Most proponents of the belief view adopt a representationalist framework[5]. They hold that believing that P requires having a mental representation with the propositional content P. However, merely having a representation with the propositional content P is not

[5] The dialectic doesn't change if one is not a representationalist. If one thinks that believing is constituted by a cluster of dispositions, the evidence reviewed in this section works for such accounts of beliefs. This evidence shows that IAs behave like beliefs, so a dispositionalist should agree that they are beliefs.

sufficient for belief. For example, one might suspect that Paris is the capital of France. One might desire that Paris be the French capital (in a scenario where Paris is demoted to the rank of an ordinary city). In philosophical jargon, desiring, suspecting and so on are called "propositional attitudes". One might have different propositional attitudes to the same content. So, propositional content alone doesn't make the mental state a belief.

What distinguishes believing from other propositional attitudes is its functional profile – that is, the way this state interacts with other states and behaviors. Differently put, the functional profile of a state is the role that this state plays in cognitive economy. For example, beliefs interact with desires in ways that other mental states don't. If one believes that Paris is the capital of France and one wants to visit the Frensh capital, one will probably buy a ticket to Paris. By contrast, if one merely suspects that Paris is the capital of France, this behavior is less likely to follow.

Although the exact specification of belief's functional profile is a matter of debate, many agree that beliefs are sensitive to evidence, inferentially interact with other personal-level mental states and play action-guiding role (Schwitzgebel 2023). For present purposes, we need not settle on an exact account. All we need is to agree on *some* of its features and then show that IAs have these features. If no competing view of the nature of IAs can account for them, the belief view becomes the best explanation of the data.

To sum up, the notion of the belief operative in the debate amounts to the following: To believe that P is to have a representation with content P that plays a belief-like functional role. So, if one wants to show that a mental state is a belief, one must examine how this state behaves. If its behavior is sufficiently similar to that of ordinary beliefs, the state is likely to be a belief.

Now we can turn to evidence for the belief view. Evidence of this kind has multiplied in recent years. A first line of support comes from studies showing that IA acquisition closely resembles

belief acquisition. But first, it is useful to contrast belief acquisition with association formation. Associations are sensitive to spaciotemporal contiguity: roughly, one associates A and B if A and B repeatedly co-occur in one's experience, like salt and pepper do. Repeated co-occurrence is the key driver of association formation.

Belief acquisition, by contrast, is sensitive to a wider range of factors. Here is a non-exhaustive list. First, like associations, beliefs respond to co-occurrences: after seeing that B follows A on many occasions, one can come to believe that A and B are somehow connected. Second, they respond to verbal instructions: telling someone that P often results in their believing that P. Third, beliefs are sensitive to relational information. Compare two situations: in one, A causes B; in the other; A makes B disappear. In both, A and B are spatiotemporally contiguous, so one might associate A and B in each scenario. That's because association formation only tracks co-occurrence between stimuli. By contrast, one will form different beliefs in the two scenarios.

Now, IAs seem to be sensitive to all these factors. Gregg et al (2006) have found that one-shot language-based learning is an effective means of producing IAs. Moreover, the same study found that such one-shot learning is *more* effective than many rounds of conditioning. This finding has been replicated in several experiments (De Houwer 2006, Kurdi and Banaji 2017). IAs have also been shown to be sensitive to relational information (see Cone et al. 2017 for a review). For example, Kurdi et al. (2022) presented participants with scenarios involving interactions between objects. Objects of a certain shape (say, blue circles) were causally responsible for positive outcomes, whereas objects of another shape (say, green triangles) simply co-occurred with such outcomes. It was found that participants developed more positive IAs when objects were causally responsible for outcomes than when they merely co-occurred with them.

A second line of evidence for the belief view concerns IA update. Again, contrasting beliefs with associations is instructive. Suppose that one wants to break the association between salt and pepper. One way to do it is to manipulate the stimuli in such a way that salt never co-occurs with pepper. The link between the concepts will eventually get weakened, and one will no longer think about one upon seeing the other. Alternatively, one might try to overwrite the association by pairing salt (or pepper) with something else. Either way, changing the association involves manipulating contingencies (cf. Mandelbaum 2016).

By contrast, belief revision is a more complex process that is not confined to just manipulating contingencies. Recent studies show that IA update is sensitive to the same factors as belief update. Cone et al. (2019) found that the source of information matters. More precisely, the information that is believed to come from a reliable source (e.g., a police report) has more impact on the existing IA than the message that is described as coming from rumors.

They also found that IA update is responsive to the believability of information. They used the setup of the already discussed study by Gregg et al (2006): participants read a story about two fictitious tribes; one tribe was portrayed positively and the other negatively. Not surprisingly, participants formed positive IAs to one and negative IAs to the other. Then they read a narrative depicting how the tribes had changed over time: the good tribe had become bad, and the bad tribe had become good. Gregg et al. (2006) found that this manipulation has no effect on IAs – a pattern that is difficult to reconcile with the belief view. However, Cone and Ferguson demonstrated that, in this setup, some participants do change their IAs, and this change is modulated by what they think about the plausibility of the narrative. Those who think that the events described in the story could have happened change their IAs more than those who find the events implausible.

Moreover, we know now that not all information is equally weighed in IA revision. To illustrate: imagine that you hear a graphic description of your colleague's bad table manners that is full of negative adjectives like "disgusting" and "repulsive". As a result, your attitude to the colleague might become slightly negative. Now, imagine that after all that, you are told that the colleague heroically saved a child from drowning. Very likely, your attitude will shift from being slightly negative to being extremely positive. That's because this single positive piece of information weighs more than the negative pieces from the bad table manners story. A series of studies looked at the factors determining the weight of a piece of information. One factor is diagnosticity: roughly, the fact that information reveals important aspects of a person's character. Diagnostic information is more successful in changing IAs than non-diagnostic information (Cone and Ferguson 2015).

Another factor has to do with whether new information allows one to re-interpret the target's previous behaviors in a new light. Mann and Ferguson (2015) have participants read a narrative about a man called "Francis West". The story revealed that Francis West broke into the home of his neighbors when they were not at home. He heavily damaged the property: he broke doors and windows and removed some "precious things" from the bedrooms. Predictably, participants formed negative IAs towards the man. But here is the twist: later they learned that Francis West did all these things because the house was on fire, and he wanted to save the children trapped inside – the words "precious things" referred to them. This new information dramatically shifted the IAs to Francis West – from extremely negative to extremely positive.

Interestingly, learning about an unrelated heroic act didn't have such a strong impact. In a follow-up study, participants read the story about Francis West without the twist and were then told that he jumped down onto subway tracks to rescue a baby from an approaching train. This manipulation didn't result in the reversal of IAs.

A final line of evidence for the belief view comes from the way IAs interact with other mental states. One important feature of belief's functional profile is that they are used as premises in reasoning. If one believes that Paris is the capital of France, this information is used in relevant inferences: for instance, if one also believes that *Mona Lisa* is kept in a museum in the capital of France, one will come to believe that *Mona Lisa* is kept in Paris[6]. By contrast, other types of mental states don't behave this way: imagining that P together with believing that if P then Q doesn't result in the belief that Q.

Like beliefs, IAs seem to be used in inferences. Mandelbaum (2016) points to a peculiar pattern in behavior of IAs: if one has a negative IA towards A and one learns that A dislikes B, one forms a positive IA towards B (Gawronski 2005). It seems that in this experiment, participants engage in the following piece of reasoning: A is a bad person; A dislikes B; so, B must be a good person. This suggests that IAs can be the premises and conclusions of an inference.

Consider one more piece of evidence pointing in this direction. Kurdi and Dunham (2020) presented participants with conditional statements about individuals belonging to two fictitious tribes. For instance, a participants might read: "If you see a blue square, you can conclude that <the target individual> is cruel". After the statement, a geometric shape appeared on the screen. In some trials, the shape matched the one mentioned in the antecedent (say, a blue square), while in others, it didn't. The conditionals and the subsequent shapes were arranged in such a way that one tribe was systematically linked to positive characteristics, while the other was connected to only negatives ones. Here is an illustration: Suppose that X belongs to Tribe A. A participant reads a series of statements attributing negative traits to X. All these statements are followed by the matching geometric shapes. Given this setup, it is rational to conclude that X

[6] There are some caveats. First, people are not logically omniscient: they don't know all the deductive consequences of their beliefs. So, the inferences one typically makes with beliefs should be restricted to some simple patterns like modus ponens. Second, beliefs interact only if they are simultaneously accessible. They might not be so accessible if one is not paying attention, tired and so.

possesses the negative traits. By contrast, all statements attributing positive traits to X are followed by non-matching shapes, so it doesn't follow that X has the relevant positive characteristics. So, the participant is expected to form negative beliefs about Tribe A. The key finding is that IA formation exhibits the described pattern. In other words, presenting a conditional (If P, then Q) and the matching shape (P) results in the IA with the content Q, whereas reading the conditional alone doesn't do so. Again, this suggests that IAs can feature in inferential processes.

Clearly, the reviewed evidence supports the belief view. But what does it mean for competing accounts? Many of the discussed studies were designed to test predictions specific to associationism, so the findings are clearly at odds with the view that IAs have associative structure. For instance, the inferential pattern described earlier – "A is bad; A dislikes B; therefore, B is good" – poses a clear problem for associationism. In this setup, B is paired with two negatively-valences ideas – DISLIKE and A, so B itself should become negatively-valenced. Or consider the Francis West studies. From the associationist perspective, the unrelated heroic act and the heroic act that reinterprets the man's behavior should produce similar IA change, since both are equally positive. Yet, only the reinterpreted story leads to a significant IA change.

What about the other accounts? Mandelbaum (2016) raises the worry that aliefs might not be that different from associations. After all, associative networks can be composed of concepts, valence and motor representations, so, like aliefs, they are also made up of representational, affective and behavioral components. Therefore, if aliefs' representational components have associative structure, aliefs are not meaningfully distinct from associations[7]. If so, the alief

[7] What if representational components are not associative but propositional? In this case, aliefs are not that different from beliefs, goes the objection raised by Mandelbaum (2016). Beliefs are made up of concepts; so, a tokening of a belief consists in, among other things, tokenings of its constituent concepts, which, again, might be associatively linked to valences and motor representations.

account faces the same problems as associationism. For instance, it can't explain IA sensitivity to the perceived reliability of the source of information, to the believability of the message, and the fact that IA feature in inferential processes. Brownstein's FTBA account closely resembles the alief view, and for this reason, it faces the same problem of not being meaningfully distinct from associationism.

Nanay's mental imagery account and Sullivan-Bissett's unconscious imaginings account are better positioned to accommodate some of the evidence discussed earlier. On their views, IAs might have propositional content. This feature can explain inferential patterns in behavior of IAs. The accounts can also explain why Francis West's unrelated heroic has less impact than the heroic casting his behaviors in a new light. The twist in West's story may allow the reader to re-imagine the whole house-breaking scene in a way that makes the protagonist look like a good guy.

But these accounts have trouble explaining other findings. For instance, it is not clear why the source of information matters for IA update. The police report and the rumor may give rise to equally vivid mental imagery, so it remains unclear why the report is more impactful. The accounts also can't accommodate the importance of believability for attitude change. Believable stories are not necessarily the ones that are easier to imagine or the ones that are imagined more vividly. Conversely, it might be easier to imagine *un*believable stories, and such stories might give rise to more vivid imaginings. These epistemic variables point more naturally to belief-like processing, where considerations of evidence quality and source reliability matter.

The evidence puts pressure on Machery's trait view, too. This view is motivated by the claim that no single mental state type can be responsible for the dispositions constitutive of implicit

bias. However, the discussed evidence suggests these dispositions are belief-like, so it is natural to conclude that the states causing them are beliefs.

Schwitzgebel's view faces the following problem. He argues that the cognitive and behavioral dispositions that constitute IAs are different from dispositions constituting ordinary beliefs. Yet, the evidence outlined here shows these dispositions may not be that different. For example, IA acquisition is sensitive to the same factors as belief acquisition, and the same is true of IA update. Given this, it becomes harder to defend the claim that IAs and beliefs are characterized by different dispositions.

Johnson's functional account is compatible with the evidence for the belief view. That's because the view is silent on what mental states play the functional role of IAs. And yet, the belief view has an explanatory advantage: it is not only consistent with the data – it can causally explain it. So, unless there is evidence that IA *don't* behave like beliefs, the belief view is in a better position than the functional account.

Overall, the belief view seems to have several advantages over the other accounts. First, it can accommodate the discussed evidence, whereas many of the other views struggle to do so.

Second, it doesn't only accommodate the evidence but also explains it. Consider IA sensitivity to what we might call "the epistemic factors" – the believability of a message and the perceived validity of a source of information. This feature makes sense on the assumption that IAs are beliefs. Arguably, one of the functions of belief is to track reality, and this function explains why mental states of this kind respond to epistemic factors. Contrast this explanation with the way the functional account deals with the given evidence. Its proponent might say something along these lines: "IAs are individuated by the role they play in cognitive economy. The evidence in question suggests that sensitivity to epistemic factors is part of IA functional profile. So, we should add this information to what we already know about this profile." While

this move allows to accommodate the evidence, it doesn't seem to explain it. The same can be said about Machery's trait view or Schwitzgebel's dispositional account.

Third, the belief view has predictive potential. This advantage is especially relevant for those who do empirical work on IAs. If we think that IAs are beliefs, we know what to expect from them because we are familiar with the way beliefs behave. This knowledge would allow us to design experiments and test the hypothesis that IAs are beliefs. Again, comparison with other accounts is instructive. For instance, a dispositionalist holds that IAs are constituted by a cluster of dispositions, but she doesn't specify what those dispositions are. Without such specification, it is impossible to formulate and test conjectures about IAs.

Finally, the belief view yields a more parsimonious ontology than some other accounts. Arguably, we should avoid positing mental exotica to explain new data if this can be done using well-tried vocabulary of beliefs, perceptual states and so on. For this reason, the belief view should be preferred to, say, the alief account, the other things being equal.

The other accounts of IAs may enjoy some of these advantages, but only the belief view seems to enjoy all of them. This makes it one of the strongest contenders in the debate.

Section 7: A challenge for the belief view

The previous section presented extensive evidence suggesting that IAs often behave like ordinary beliefs. Over the past decade, findings of this sort have accumulated to the point where it is difficult to deny that IAs are somewhat belief-like. But can proponents of the belief view declare victory?

This reaction is premature because, while IAs often behave like ordinary beliefs, they diverge from them in important respects. Here is a template of a recurring objection along these lines:

IAs have feature F. Ordinary beliefs don't have feature F. Therefore, IAs can't be ordinary beliefs. The more such features one identifies, the stronger the objection becomes. If this challenge is not addressed, a proponent of the belief view is pushed towards a position like Levy's: even though IAs share many features with beliefs, they differ from in significant ways. Hence, they belong to a peculiar kind of mental state for which we currently don't have a name. Alternatively, this situation may be taken to support a heterogeneity thesis in the spirit of Holroyd and Sweetman (2016): the fact that IAs exhibit belief-like and non-belief-like patterns suggests that such attitudes are constituted by several *types* of mental states.

What features distinguish IAs from ordinary beliefs? One frequently cited difference concerns attitude update (Madva 2016). IAs do not appear to be sensitive to all the factors that influence belief revision. For one thing, IAs often fail to respond to negation. In one study, participants were presented with stereotype-congruent pairs of stimuli (say, a female name followed by the word "weakness") and instructed to respond "no". Surprisingly, this did not weaken the IA in question – it strengthened them (Gawronski et al. 2008). For another, IAs fail to take into account some crucial information about stimuli. In another study, participants read descriptions of novel social targets, and later they were told the descriptions were false. However, this information had little impact on the IAs (Peters and Gawronski 2011).

The belief view has resources to accommodate update failures. One can argue that such failures are retrieval-related: roughly, a belief is not updated by given evidence because the belief is not retrieved when the evidence is presented (De Houwer et al. 2019). Of course, this proposal requires further elaboration – at the very least, one should specify the factors that modulate retrieval.

The fragmentation approach to belief storage often such elaboration. This approach is motivated by the observation that people sometimes hold blatantly inconsistent beliefs. For

example, one can believe that one has a dentist appointment on Monday at 10 am, and at the same time, one can believe that one's Monday morning is free. The inconsistency is obvious, and yet one may not notice it. The explanation, according to the fragmentation approach, is that these beliefs are stored in different fragments or "folders". A belief is retrieved only when the fragment storing it is activated. If two plainly inconsistent beliefs are kept in different fragments, the inconsistency may go unnoticed.

Bendana and Mandelbaum (2020) apply this framework to IAs. On their account, the IAs about the novel social targets described in Peters and Gawronski (2011) are unaffected by the corrective information because the fragment storing the IAs remains inactive when the new information is presented. Bendana and Mandelbaum (2020) propose several principles for fragment creation, maintenance and activation. For example, they speculate that entering a novel environment might result in creating a new fragment. It remains to be seen whether the principles can explain the failures of update described in IA literature.

In this section, however, I focus on another difference between IAs and ordinary beliefs – one that concerns self-knowledge. A standard view holds that people are typically in a good epistemic position with respect to their beliefs: if one believes that P, one is usually in a position to know (or at least to have the justified belief) that one believes that P. By contrast, empirical work on IAs suggests that this is not true of IAs. People are often not aware of their IAs. But if IAs are beliefs, how could this be the case? Call this objection "the self-knowledge objection" to the belief view.

In the rest of this section, I will briefly examine this objection. The full-blown discussion can be found in Chapter 2.

Why think that one's epistemic position vis-à-vis IAs is poor? Let's call this feature of IAs "opaqueness". There are several reasons to think that IAs are opaque. First, IAs were

discovered by using indirect measures, such the IAT or the AMP, which don't rely on self-report. Why doesn't self-report capture them? Presumably because IAs are not available for introspection.

Second, there is empirical evidence that people are often surprised by their results on indirect tests of attitudes. For example, one study has found that a significant number of participants were stunned by their IAT scores (Monteith et al. 2001). This finding is hard to explain unless one thinks that IAs are opaque.

Third, the inconsistency between EAs and IAs should give rise to the feeling of cognitive dissonance. Literature on cognitive dissonance shows that this feeling prompts one to eliminate the inconsistency (Cooper 2007). The fact that it doesn't happen in the case of IAs and EAs suggests that one is not aware of the inconsistency between the attitudes. Again, the opaqueness of IAs provides an explanation (Berger 2020).

That said, opaqueness does not imply complete inaccessibility. It doesn't mean that one never has access to one's IA. Some studies show that people can accurately predict their IATs results (Hahn et al. 2014). However, a closer look at the experimental setup reveals that participants might be inferring their IATs from certain cues, such as their affective reactions to certain stimuli. If so, IAs seem to be more opaque than ordinary beliefs.

Why can't we just say that IAs are unconscious beliefs, whereas ordinary beliefs are conscious? This would quickly address the self-knowledge objection. One problem with this reply is that conscious mental states have phenomenal character, and beliefs, arguably, lack such character. If this is right, *all* beliefs are unconscious, and it would make no sense to talk about conscious beliefs. So, one wouldn't be able to use the conscious/unconscious distinction to capture the difference between IAs and ordinary beliefs.

The other problem is that, even if we grant that IAs are unconscious and ordinary beliefs are conscious, we would still need an explanation of why some beliefs are unconscious, whereas other beliefs are conscious. Just stating that IAs are unconscious doesn't solve the problem but pushes it one step further.

My reply to the self-knowledge objection amounts to arguing that IAs constitute a special kind of beliefs – beliefs that are not accompanied by the relevant judgments. More precisely, I argue that when one implicitly believes that P, a tokening of this belief is not accompanied by judging that P. By contrast, tokenings of ordinary beliefs are usually accompanied by such judgments. By "judging", I mean entertaining a proposition and endorsing it.

Now, judgment is a common route to self-knowledge of belief. If one wants to know whether one believes that P, one typically asks whether P is the case (Evans 1982). If the available evidence supports P, one tends to judge that P and thus concludes that one believes that P. In other words, one often knows what one believes by looking at one's judgments. IAs are not accompanied by the relevant judgments, so this route to self-knowledge is blocked for them. This explains their opaqueness.

Why aren't IAs accompanied by the relevant judgments? I speculate that IAs might akin to what we might call "superstitious beliefs", like the belief that spilling salt brings bad luck. Such beliefs are evidentially problematic by the agent's own lights. As a result, they are not aligned with the relevant judgments.

This account of IAs raises the following worry: if I explain access to ordinary beliefs via the link between believing and judging, but deny that IAs are accompanied by judgments, doesn't this undermine the claim that IAs are beliefs? To address this worry, I need to provide an account of belief that explains why beliefs tend to be accompanied by judgments and, at the same time, allows for the fact that some beliefs are *not* accompanied by judgements.

An appeal to teleology and certain features of cognitive architecture might do the trick. Beliefs have the function of carrying information about the world, and to fulfil this function, they must be responsive to evidence. This explains why beliefs are typically aligned with evidence and judgments. At the same time, belief fixation might be ballistic: a tokening of any truth-apt mental representation P automatically yields believing P (Gilbert 1990, Mandelbaum 2013). Moreover, belief storage might be fragmented, as discussed earlier. These constraints explain how a state with the function of tracking reality often fails to fulfill it.

This applies to IAs. They are beliefs, they belong to the mental state type whose function consists in tracking reality, and yet they fail to carry out this function. This results in an interesting combination: on the one hand, IAs feature in belief-like processing; on the other hand, they are not endorsed by the agent because they are epistemically problematic by her own lights. In this respect, they may differ from the majority of one's beliefs. Nonetheless, this doesn't prevent them from being beliefs. As we saw in previous sections, IA behavior is still very similar to that of ordinary beliefs.

Section 8: Two approaches to inference

As we saw before, some evidence for the belief view relies on the idea that beliefs feature in inferential transitions. For example, if one has a negative IA towards A and one learns that A dislikes B, one forms a positive IA towards B (Gawronski 2005). It looks like the resulting IA towards B is the output of the following inference: A is a bad person; A dislikes B; so, B must be a good person. This in turn suggests that the attitude towards B is a belief. The argument can be formulated as follows: If a mental state features in an inferential process as a premise or a conclusion, this state is a belief. This reasoning fits well with a widely held view that beliefs are the kind of mental states that are used as premises in reasoning.

To assess the argument, we must clarify what inference is. This question will be addressed in detail in Chapter 3. In this section, I will outline the dialectic that frames that chapter. I will argue that there are two approaches to understanding inference and that only one of them supports the argument just presented: If a state features in an inference, this state is a belief. The other approach is compatible with associative states featuring in inferential transitions.

Here is an example to pump the intuition that associations feature in inference. Imagine a detective with an excellent visual memory. She has spent many hours examining the database of fingerprints in the police department. This database contains pictures of fingerprints paired with the names of the individuals to whom the fingerprints belong. One day, the detective arrives at a crime scene. The forensic expert shows her the fingerprints of the suspect. The detective examines them and declares: "They belong to John Smith!"

How should we describe the detective's transition from perceiving the fingerprints to believing that they belong to John Smith? Some would say that the detective *inferred* the suspect's identity from the evidence. But here is the crucial detail of the story: associative links played an important role in the transition. The detective's examination of the fingerprint database has led to the formation of associative links between the fingerprints and the names. As a result, when the detective sees a fingerprint, the relevant name comes to her mind.

Does the associative basis of the transition undermine its status as an inference? Should one now say that this transition is *not* inferential because it was underwritten by associative links? Here, intuitions differ. Some are inclined to say that inference and association are mutually exclusive notions. Others would argue that the two are compatible.

This disagreement illustrates the difference between the two approaches to inference. On one view, which I will call *"*the mental-processes approach*"*, inference is defined in terms of the psychological structure underwriting transitions between mental states. For instance, Quilty-

Dunn and Mandelbaum (2017) that a transition count as inferential if and only if it is produced by mental processes sensitive to the logical form of the representations involved. By contrast, the other approach – let's call it "epistemic" – spells out the notion of inferential transitions in epistemic terms while abstracting away from the structure of mental processes underwriting them. Siegel (2019) provides an account of inference exemplifying this approach. It holds that inferring is a special kind responding to mental states in which the output mental state epistemically depends on the input.

Chapter 3 examines Siegel's account in detail. I clarify her view by offering the following interpretation. A transition from state A to state B is inferential if and only if A causes B and B is epistemically dependent on A. Intuitively, B epistemically depends on A just in case A's justificatory status is relevant B's justificatory status. I suggest that this relevance can be tested by the following counterfactuals:

(C) If A were not justified, B wouldn't be justified.

(C*) If A were justified, B would be justified.

I then argue that this epistemic account is compatible with some associative transitions being inferential. To show this, I describe several scenarios similar to the detective example. Interestingly, Siegel herself thinks that inference and association are mutually exclusive. For this reason, the scenarios constitute counterexamples to her account. I suggest that the most natural way to accommodate these cases is to abandon the claim about the incompatibility of inference and association. This claim doesn't follow from the core of her account and can be dropped without theoretical loss.

Returning to IAs, we can now reconsider the argument: IAs feature in inferential transitions; therefore, IAs are beliefs. It works only if one adopts the mental-processes approach to inference. On this approach, what makes a transition inferential are certain psychological

properties of the mental states and processes involved. If it turns out that only beliefs have these properties, the argument is valid. However, if one goes for the epistemic approach, the reasoning is problematic. As we have just seen, this approach allows for the possibility that states with associative structure feature in inferential transitions. Accordingly, the fact that IAs participate in inference-like patterns does not entail that they are beliefs. The upshot is that the force of the inference-based argument for the belief view depends crucially on which of the two approaches one adopts. Drawing attention to this fact helps to avoid futile terminological debates and strengthens the conceptual foundations of the debate over IAs.

References

Bendana, J. & Mandelbaum, E. (2021) *The fragmentation of belief,* in Borgoni, C., Kindermann, D., & Onofri, A. (eds.) *The fragmented mind*, Oxford University Press, pp. 78-108.

Berger, J. (2020) Implicit attitudes and awareness. *Synthese*, *197*(3), pp. 1291–1312. https://doi.org/10.1007/s11229-018-1754-3.

Brownstein, M. (2018) *The Implicit Mind: Cognitive Architecture, the Self, and Ethics*, New York: Oxford University Press.

Carruthers, P. (2018). Implicit versus explicit attitudes: Differing manifestations of the same representational structures? *Review of Philosophy and Psychology, 9*(1), 51–72. https://doi.org/10.1007/s13164-017-0354-3

Cone, J., & Ferguson, M. J. (2015). He did what? The role of diagnosticity in revising implicit evaluations. *Journal of Personality and Social Psychology, 108*(1), 37–57. https://doi.org/10.1037/pspa0000014

Cone, J., Mann, T. C., & Ferguson, M. J. (2017) Changing our implicit minds: How, when, and why implicit evaluations can be rapidly revised, *Advances in Experimental Social Psychology*, pp. 131–199. https://doi.org/10.1016/bs.aesp.2017.03.001

J. Cone, K. Flaharty, & M.J. Ferguson, Believability of evidence matters for correcting social impressions, Proc. Natl. Acad. Sci. U.S.A. 116 (20) 9802-9807, https://doi.org/10.1073/pnas.1903222116 (2019).

Cooper, J. (2007) *Cognitive dissonance: Fifty Years of a classic theory*, Sage Publications.

De Houwer, J. (2006). Using the implicit association test does not rule out an impact of conscious propositional knowledge on evaluative conditioning. Learning and Motivation, 37, 176–187 https://doi.org/10.1016/j.lmot.2005.12.002

De Houwer, J. (2014) A propositional model of implicit evaluation, *Social and Personality Psychology Compass*, *8*(7), pp. 342–353. https://doi.org/10.1111/spc3.12111

De Houwer, J., Van Dessel, P., & Moran, T. (2020). Attitudes beyond associations: On the role of propositional representations in stimulus evaluation. In B. Gawronski (Ed.), *Advances in experimental social psychology* (pp. 127–183). Elsevier Academic Press. https://doi.org/10.1016/bs.aesp.2019.09.004

Evans, G. (1982) *The Varieties of Reference*, Oxford University Press.

Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986) On the automatic activation of attitudes. *Journal of Personality and Social Psychology*, *50*(2), pp. 229–238. https://doi.org/10.1037/0022-3514.50.2.229.

Fazio, R. H. (1995). Attitudes as object-evaluation associations: Determinants, consequences, and correlates of attitude accessibility. In R. E. Petty & J. A. Krosnick (Eds.), *Attitude strength: Antecedents and consequences* (pp. 247–282).

Gawronski, B. & Bodenhausen, G.V. (2006) Associative and Propositional Processes in Evaluation: An Integrative Review of Implicit and Explicit Attitude Change, *Psychological Bulletin*, 132(5), pp. 692–731.

Gawronski, B., Deutsch, R., Mbirkou, S., Seibt, B., & Strack, F. (2008) When "just say no" is not enough: Affirmation versus negation training and the reduction of automatic stereotype activation, *Journal of Experimental Social Psychology, 44*(2), pp. 370–377.

Gawronski, B., Walther, E., and Blank, H. (2005). Cognitive Consistency and the Formation of Interpersonal Attitudes: Cognitive Balance Affects the Encoding of Social Information. Journal of Experimental Social Psychology, 41, 618–626.

Gendler, T. S. (2008) Alief and belief, *Journal of Philosophy* 105(10), pp. 634-663.

Gilbert, D. (1991) How mental systems believe. American Psychologist, 46(2), 107–119. https://doi.org/10.1037//0003-066x.46.2.107

Gregg, A., Seibt, B. & Banaji, M. (2006) Easier Done than Undone: Asymmetry in The Malleability of Implicit Preferences, *Journal of Personality and Social Psychology* 90 (1), pp. 1-20

Greenwald, A. G. & Banaji, M. R. (1995) Implicit social cognition: Attitudes, self-esteem, and stereotypes, *Psychological Review*, 102(1), pp. 4–27.

Greenwald, A., M. Banaji, & B. Nosek. (2015) Statistically small effects of the implicit association test can have societally large effects, *Journal of Personality and Social Psychology* 108, pp. 553–561.

Greenwald, A. G. & Lai, C. K. (2020) Implicit social cognition, *Annual Review of Psychology*, *71*(1), pp. 419–445. https://doi.org/10.1146/annurev-psych-010419-050837.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998) Measuring individual differences in implicit cognition: The Implicit Association Test, *Journal of Personality and Social Psychology*, *74*(6), pp. 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464.

Hahn, A., Judd, C. M., Hirsh, H. K., & Blair, I. V. (2014) Awareness of implicit attitudes, *Journal of Experimental Psychology: General*, 143(3), pp. 1369–1392. https://doi.org/10.1037/a003502.

Holroyd, J. (2016). VIII—what do we want from a model of implicit cognition? *Proceedings of the Aristotelian Society*, *116*(2), pp. 153–179. https://doi.org/10.1093/arisoc/aow005

Holroyd, J. & Sweetman, J. (2016) The heterogeneity of implicit bias. *Implicit Bias and Philosophy, Volume 1*, pp. 80–103. https://doi.org/10.1093/acprof:oso/9780198713241.003.0004

Johnson, G.M. (2020) The Structure of Bias, *Mind* 129/516, pp. 1193–236.

Johnson, I. R., Kopp, B. M., & Petty, R. E. (2016) Just say no! (and mean it): Meaningful negation as a tool to modify automatic racial attitudes, *Group Processes & Intergroup Relations*, *21*(1), pp. 88–110. https://doi.org/10.1177/1368430216647189

Kurdi, B., & Banaji, M. R. (2017). Repeated evaluative pairings and evaluative statements: How effectively do they shift implicit attitudes? Journal of Experimental Psychology: General, 146, 194–213.

Kurdi, B. & Dunham, Y. (2020) Propositional accounts of implicit evaluation: Taking stock and looking ahead. Social Cognition, 38 (Suppl), pp. 42-67. https://doi.org/10.1521/soco.2020.38.supp.s42

Kurdi, B., & Dunham, Y. (2021). Sensitivity of implicit evaluations to accurate and erroneous propositional inferences. *Cognition*, 214, Article 104792. https://doi.org/10.1016/j.cognition.2021.104792

Kurdi, B., Morris, A., & Cushman, F. A. (2022). The role of causal structure in implicit evaluation. *Cognition, 225,* 1–19. https://doi.org/10.1016/j.cognition.2022.105116

Levy, N. (2014). Consciousness, implicit attitudes and moral responsibility. Nous 48(1): 21–40. https://doi.org/10.1111/j.1468-0068.2011.00853.x.

Levy, N. (2015) Neither fish nor fowl: Implicit attitudes as patchy endorsements, *Noûs*, 49(4), pp. 800–823. https://doi.org/10.1111/nous.12074

Machery, E. (2016) De-Freuding Implicit Attitudes, in Brownstein, M., & Saul J. (eds.) *Implicit bias and Philosophy*, vol. 1, pp. 104-129

Madva, A. (2016) Why implicit attitudes are (probably) not beliefs, *Synthese*, *193*(8), pp. 2659–2684. https://doi.org/10.1007/s11229-015-0874-2

Mandelbaum, E. (2013) Thinking is believing, *Inquiry*, *57*(1), 55–96. https://doi.org/10.1080/0020174x.2014.858417

Schwitzgebel, E. (2023) *Belief*, [Online], Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/entries/belief/ [27 January 2025]

Mandelbaum, E. (2022) *Associationist Theories of Thought* [Online], Stanford Encyclopedia of Philosophy https://plato.stanford.edu/archives/win2022/entries/associationist-thought/ [27 April 2025]

Mandelbaum, E. (2016) Attitude, association, and inference: On the propositional structure of implicit bias, *Noûs* 50(3), pp. 629-658

Mann, T. C., & Ferguson, M. J. (2015). Can we undo our first impressions? The role of reinterpretation in reversing implicit evaluations. Journal of Personality and Social Psychology, 108(6), 823–849. http://doi.org/10.1037/pspa0000021.

Mann, T. C., Kurdi, B., & Banaji, M. R. (2020) How effectively can implicit evaluations be updated? using evaluative statements after aversive repeated evaluative pairings. *Journal of Experimental Psychology: General*, *149*(6), pp. 1169–1192. https://doi.org/10.1037/xge0000701.

McConnell, A. R., & Rydell, R. J. (2014). The systems of evaluation model: A dual-systems approach to attitudes. In J. Sherman, B. Gawronski, & Y. Trope (Eds.), Dual process theories of the social mind (pp. 204–217). New York: Guilford.

Monteith, M. J., Voils, C. I., & Ashburn-Nardo, L. (2001) Taking a look underground: Detecting, interpreting and reacting to implicit racial biases, *Social Cognition* 19(4), pp. 395–417.

Nanay, B. (2021) Implicit bias as mental imagery, *Journal of the American Philosophical Association*, *7*(3), pp. 329–347. https://doi.org/10.1017/apa.2020.29.

Nosek BA, Banaji MR. (2001) The go/no-go association task, Social Cognition 19(6), pp. 625–66.

Oswald, F., Mitchell G., Blanton, H., Jaccard J., & Tetlock, P. (2013) Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies, *Journal of Personality and Social Psychology* 105, pp. 171– 192.

Papineau, D. (2013) *There are no norms of belief,* in T. Chan (ed.) *The Aim of Belief*, Oxford University Press, pp. 64–79.

Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, B. D. (2005) An inkblot for attitudes: Affect misattribution as implicit measurement, *Journal of Personality and Social Psychology*, *89*(3), pp. 277–293. https://doi.org/10.1037/0022-3514.89.3.277.

Peters, K. R., & Gawronski, B. (2011). Are we puppets on a string? Comparing the impact of contingency and validity on implicit and explicit evaluations. *Personality and Social Psychology Bulletin*, *37*(4), 557–569. https://doi.org/10.1177/0146 167211400423

Petty, R. E., Brin˜ol, P., & DeMarree, K. G. (2007). The meta-cognitive model (MCM) of evaluations: Implications for evaluation measurement, change, and strength. Social Cognition, 25, 657–686.

Quilty-Dunn, J., & Mandelbaum, E. (2017). Inferential transitions. Australasian Journal of Philosophy, 96(3), 532–547. https://doi.org/10.1080/00048402.2017.1358754

Siegel, S. (2019). Inference without Reckoning. in Reasoning: Essays on Theoretical and Practical Thinking, ed. M. Balcerak-Jackson and B. Balcerak-Jackson, Oxford: Oxford University Press.

Schwitzgebel, E. (2002). A Phenomenal, Dispositional Account of Belief. Noûs 36 (2):249-275.

Schwitzgebel, E. (2023) *Belief*, [Online], Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/entries/belief/ [10 September 2024]

Sullivan-Bissett, E. (2019) Biased by Our Imaginings, *Mind and Language* 34/5, pp. 627–47.

Toribio, J. (2018) Implicit bias: From social structure to representational format. *THEORIA. An International Journal for Theory, History and Foundations of Science*, *33*(1), pp. 41-60. https://doi.org/10.1387/theoria.17751

Chapter 2: Beliefs without judgments: a plea for the belief view of implicit attitudes

Article published in *Journal of Consciousness Studies,* **32**, No. 1–2, February 2025, pp. 160-185.

https://doi.org/10.53765/20512201.32.1.165

Abstract

Implicit attitudes (IAs) are mental states that are responsible for discriminatory behavior called "implicit bias". There is no agreement about the nature of IAs. Some argue that they don't differ from beliefs. The paper defends this view from the following objection: One is in a good epistemic position with respect to one's beliefs: if one believes that P, one tends to know that one believes that P. However, studies show that often people are not aware of having IAs. How can it be if IAs are beliefs? I address this objection by defending the claim that implicit beliefs constitute a special kind of belief – beliefs that are not accompanied by the relevant judgments. More precisely, if one implicitly believes that P, a tokening of this belief is not accompanied by judging that P. Since judging is a route to knowing one's beliefs, this route is blocked for implicit beliefs.

Keywords: belief, implicit attitudes, judgment, unconscious belief, self-knowledge, transparency

Introduction

The development of indirect measures[8] of attitudes has made it clear that explicit non-egalitarian beliefs are not the only source of discriminatory behavior. People who sincerely profess egalitarian beliefs often behave in ways that are not consistent with such beliefs. We know now that the reason for this discrepancy is that people not only have explicit attitudes (EAs) towards a given social group – they also have implicit attitudes (IAs), which are often at odds with their explicit counterparts[9]. The existence of such divergent attitudes allows for the possibility that committed egalitarians might routinely be engaging in behaviors that they themselves find reprehensible and wrong. This phenomenon gives rise to a number of puzzling questions: What is the nature of the mental states underwriting these discriminatory behaviors that go against one's beliefs and values? Is the agent morally responsible for such actions? Is the agent aware of them and the mental states that drive them?

The paper explores a different puzzle emerging from the discovery of IAs. The view holding that IAs are beliefs has recently been gaining popularity among philosophers and psychologists (De Houwer 2014, Frankish 2016, Mandelbaum 2016). On this view – call it "the belief view" – mental states captured by indirect measures are not different in kind from ordinary beliefs, like the belief that Paris is the capital of France. The belief view enjoys serious empirical support, but it faces an objection stemming from the fact that IAs don't seem to behave like ordinary beliefs do. The more differences there are between IAs and ordinary beliefs, the more

[8] The Implicit Association Test (Greenwald et al. 1998), evaluative priming (Fazio et al. 1986), affect misattribution procedure (Payne et al. 2005), Go/No-Go Association test (Nosek and Banaji 2001), to name a few. Unlike more traditional measures of attitudes like questionnaires or feeling thermometers, these tests don't rely on self-report – instead, one's attitude is inferred from some features of one's test performance, such as response latencies. For a recent review of indirect measures, see Greenwald and Lai 2020.

[9] I use "implicit attitudes" to refer to mental states causing certain discriminatory behaviors. This usage of the term is not idiosyncratic – see Mandelbaum 2016, Nanay 2021, Karlan 2021. However, a few authors prefer different terminology and use the expression "implicit bias" to that end (Holroyd and Sweetman 2016, Toribio 2018, Sullivan-Bissett 2019).

pressing this objection is. In this paper, I focus on one such difference[10]. On the one hand, one tends to be in a good epistemic position with respect to one's beliefs. If one believes that P, one is in a position to know (or at least to have the justified belief) that one believes that P. By contrast, empirical studies suggest that one's epistemic position vis-à-vis IAs is poor, since often people are not aware of having IAs (Berger 2020). How can it be the case, the objection goes, if IAs don't differ from ordinary beliefs? Call this objection "the self-knowledge objection" to the belief view[11].

The goal of the paper is to address the self-knowledge objection and thus provide indirect support for the belief view. I argue, in a nutshell, that implicit beliefs constitute a special kind of belief[12]: if one implicitly believes that P, a tokening of this belief is not accompanied by judging that P. This feature of implicit beliefs explains their limited self-knowledge because judgment is an important route to self-knowledge. The latter claim is supported by Evans's observation that beliefs are transparent (Evans 1982): if one wants to know whether one believes that P, one must look not at the mental state in question (the belief that P), but through it (hence, the metaphor of transparency) – at the state of affairs that it represents, i.e., P.

I will proceed as follows. In section 1, I introduce the belief view of IAs and outline some evidence in its favor. Section 2 motivates the claim that one's epistemic position vis-à-vis IAs is poor. I call this feature of IAs "opaqueness." Section 3 discusses one way of addressing the self-knowledge objection, which holds that implicit beliefs are unconscious. I give some reasons for being skeptical about this solution. In section 4, I outline my account of implicit

[10] Another difference has to do with IAs' failure to update in light of relevant evidence (Gregg et al. 2006). This patchy responsiveness to evidence is highlighted in Levy 2015. For an explanation of how the belief view can accommodate this feature of IAs, see Karlan 2021.

[11] I will be assuming that one's good epistemic position to one's beliefs amounts to self-knowledge. But the argument won't be affected if we replace "self-knowledge" with "one's justified belief about one's mental state" or other expressions denoting a positive epistemic status.

[12] Since implicit beliefs are a special kind of *belief*, my account counts as a version of the belief view. After all, implicit beliefs and ordinary beliefs share some important features: both interact with desires and other motivational states to give rise to behavior, and both interact with other cognitive states, giving rise to new beliefs, decisions, etc. – a feature known as 'inferential promiscuity' (Stich 1978).

beliefs as beliefs that are not accompanied by the relevant judgments. Section 5 shows how this account explains one's poor epistemic position with respect to implicit beliefs. Section 6 tackles one objection to my account of implicit beliefs by elaborating on the relationship between belief, judgment and evidence-responsiveness. Finally, section 7 discusses further advantages of my account of implicit beliefs.

Section 1: The belief view of implicit attitudes

Consider the following case:

Sam is a hiring manager in a company. Part of his job consists in going through dossiers of different candidates and deciding which of them is fit for the position. Sam sincerely endorses the claim that one's race is irrelevant to one's professional qualification. But if we take a closer look at Sam's hiring decisions, we will find out that they are not always consistent with his egalitarian views: it turns out that the dossiers of people of race R end up in the pile "Rejected" much more often than the dossiers of other candidates do, even though R's candidates are equally suitable for the job.

Sam's case is intended to be an illustration of a real-world phenomenon. Research on implicit bias provides reasons to think that people who sincerely hold egalitarian views might nonetheless engage in various subtle forms of discrimination[13]. Think about Sam. Let P stand for the claim that R's candidates are less suitable for a given job than other people. What does Sam believe – P or not-P?

[13] There is an ongoing debate about the real-world effects of IAs. See Oswald et al. 2013, for the claim that such effects are minor, and Greenwald et al. 2015 for a reply to this critique.

It is relatively uncontroversial to claim that Sam has the egalitarian belief, not-P[14]. By contrast, there is little consensus about how to describe his IA towards R. It has been proposed (to give a non-exhaustive list) that IAs are associations (Gawronski and Bodenhausen 2006, Toribio 2018), aliefs (Gendler 2008), affective tension clusters (Brownstein 2018), beliefs (De Houwer 2014, Mandelbaum 2016, Carruthers 2017), character traits (Machery 2016), mental imagery (Nanay 2021), imaginings (Sullivan-Bissett 2019), patchy endorsements (Levy 2015), functionally individuated cognitive structures that underwrite social-kind inductions (Johnson 2020).

The aim of the paper is to defend the belief view from the self-knowledge objection, so the main argument clearly presupposes this account of IAs. Due to space limitations, I can't examine in detail the evidence supporting it. Instead, I will say what the view amounts to, and outline some considerations that have been put forward in its favor. The goal of this exposition is not to conclusively show that the belief view is correct, but rather to demonstrate that, at the very least, this account is a reasonable option. For this reason, it is a worthwhile project to think about how to defend it from potential objections, such as the self-knowledge objection. Differently put, the project of this paper is a conditional one: it asks the reader to assume that IAs are beliefs and then consider whether this claim can be squared with the fact that IAs differ from ordinary beliefs in some important respects. If we show that it can, it will provide additional support for the belief view.

The belief view holds that IAs are not different from ordinary beliefs, like the belief that Paris is the capital of France. As for the Sam case, the account holds that his implicit belief P is the same kind of mental state as his explicit, egalitarian belief not-P. What does it mean to believe that P? If one is a representationalist about belief (as are most of those who defend the belief

[14] For example, Carruthers (2009) disagrees and argues that in cases like Sam's, the agent believes that she believes that not-P. In other words, she has a second-order belief about a first-order belief.

view), believing that P consists in having a representation with the content P stored in the belief box. "The belief box" stands for whatever functional role that is played by beliefs[15]. The exact functional role is a matter of controversy, but many agree that beliefs are sensitive to evidence, inferentially interact with other personal-level mental states and play action-guiding role (Schwitzgebel 2023).

A number of studies have found that IAs often behave as beliefs do, and this evidence is used to argue for the belief view. Here are some of the hallmarks of belief that IAs have been argued to have: sensitivity to the logical structure of information (Kurdi and Dunham 2021), inferential interactions with other personal-level states (Mandelbaum 2016), swift update in light of diagnostic information (Mann et al. 2020), differential sensitivity to strong vs weak arguments (Briñol et al. 2009), one-shot language-based learning (Gregg et al. 2006), sensitivity to relational information (Kurdi and Dunham 2020)[16]. To illustrate: several studies have shown that the update of IAs depends on, as it were, the quality of the information presented: information marked as a rumor leads to a weaker attitude change than information coming from what is described as a reliable source (Cone et al. 2017). Beliefs are expected to be sensitive to this factor, whereas other types of mental states, such as associations or mental imagery are not. So, the argument goes, the belief view provides the best explanation of the data[17].

A caveat: these considerations supporting the belief view don't necessarily show that the other accounts of IAs are incorrect. It is possible that IAs are a heterogeneous family of mental states:

[15] One doesn't have to be a representationalist to defend the belief view. If one holds that believing merely involves certain dispositions, IAs qualify as beliefs because they are associated with these dispositions.

[16] Some of this evidence has been challenged: IAs' inferential interactions are disputed by Levy 2015, Madva 2016 argues against IAs' sensitivity to logical form.

[17] Undoubtedly, there are data that the belief view struggles to explain. One problem is to accommodate IAs' selective recalcitrance to evidence (Gregg et al. 2006). Another is IAs' insensitivity to certain logical operators. Madva 2016 points out that IAs are not sensitive to negation and conditional. Gawronski et al. 2008 makes the same point about negation. However, evidence on negation is mixed (Boucher and Rydell 2012). And some studies indicate that negation alters IAs in certain circumstances (Johnson et al. 2016). Finally, there is recent evidence that IAs are in fact sensitive to conditionals (Kurdi and Dunham 2021).

say, some of them are beliefs, whereas others are states with associative structure[18]. Nonetheless, the evidence put forward for the belief view strongly suggests that *some* IAs are beliefs because this claim provides the best explanation for this evidence. The self-knowledge objection is thus applicable to *those* IAs, which, if the heterogeneity claims is correct, constitute a subset of all IAs. So, the argument of the paper can get off the ground.

Let me close this section with some terminological clarifications. There are many kinds of beliefs: explicit, implicit, occurrent, dispositional; moreover, each of these notions is open to different interpretations. I will adopt the following usage throughout the paper. When I talk about explicit and implicit beliefs, I refer to beliefs that underwrite explicit and implicit attitudes, respectively. Sam's belief that P is thus implicit, whereas his belief not-P is explicit. By "occurrent beliefs" I will mean beliefs that are operative: one's belief that *Mona Lisa is kept in the Louvre Museum* is operative – and therefore occurrent – when, say, one is planning one's trip to see this famous painting. A belief is dispositional if and only if it is not occurrent. These distinctions will become relevant later[19].

Section 2: The opaqueness of implicit attitudes

Recall that the self-knowledge objection to the belief view goes as follows: One tends to be in a good epistemic position with respect to one's beliefs; by contrast, one's epistemic position vis-à-vis IAs is poor. How can it be the case if IAs are beliefs? In this section, I will review

[18] Holroyd and Sweetman 2016 speculate that IAs are heterogeneous in at least two respects. First, they are constituted by associations with different types of content. Second, IAs feature in different cognitive processes. I suggest that we might need to go further and add another dimension of heterogeneity: perhaps IAs are underwritten by different *kinds* of mental states – beliefs, associations, etc.

[19] Notice that this way of cashing out the difference between occurrent and dispositional beliefs doesn't presuppose any particular account of belief – different accounts can accommodate it. A representationalist can identify an occurrent belief with a mental representation that is retrieved from memory and is poised to guide behavior and cognition. A dispositionalist can say that a belief is occurrent if and only if the dispositions constitutive of this belief are being manifested.

evidence for one's poor epistemic position with respect to implicit beliefs – the feature that we might call "opaqueness."

First of all, recall that IAs are tapped into by indirect measures: instead of asking the subject what she thinks about the target race R, the experimenter infers the attitude towards R from her performance on some task. One rationale behind introducing indirect measures was precisely that the subject might not be aware of the mental states in question[20], which would render self-report measures inefficient (Greenwald and Banaji 1995)[21]. So, the very fact that IAs are assessed this way suggests that these mental states are opaque.

Second, there is some evidence that people tend to be surprised by their results on indirect measures of attitudes. For example, one study has found that a significant number of participants were stunned by their performance on the Implicit Association Test (IAT), a widely used indirect measure of IAs (Monteith et al. 2001). More precisely, these participants noticed a discrepancy between their performance and the way they thought they should have performed, and they were surprised by this fact. A natural explanation of these results holds that participants are stunned because the mental states responsible for the performance on the IAT are difficult to detect in normal circumstances. For this reason, they are not aware of having these states and so don't expect them to have an impact on the test performance[22].

Third, weak correlations between direct and indirect measures of attitudes suggest that IAs diverge from their explicit counterparts (Greenwald and Lai 2020). Indeed, this finding is

[20] Lack of awareness is related to opaqueness in the following way: A good epistemic position vis-à-vis a mental state implies that one is in a position to form a justified belief (or a belief amounting to knowledge) about this state. However, people who are not aware of their IAs often don't have any beliefs about them, let alone justified beliefs or beliefs amounting to knowledge. This shows that lack of awareness is a good indicator of being in a bad epistemic position vis-à-vis a given state.

[21] Another reason for introducing indirect measures is reputation management: one might not want to reveal one's attitude if this attitude goes against the prevailing social norms. However, it was demonstrated that explicit and implicit attitudes diverge even in situations when there is no social pressure to hide one's IA (Wilson et al. 2000).

[22] Alternatively, one can argue that one can detect these states but has no idea of what kind of impact they have on behavior (Gawronski et al. 2006).

reflected in the Sam example: he endorses egalitarian views and at the same time implicitly believes that people of race R are less competent than other people. Now, a vast literature on cognitive dissonance teaches us that the awareness of inconsistent cognitions literally hurts: such inconsistencies give rise to the unpleasant feeling of dissonance, which prompts the agent to do something to reduce it (Cooper 2007). One way to do so is to eliminate the inconsistency that causes the feeling. So, if people with divergent attitudes experienced dissonance, they would be expected to eventually get their attitudes aligned and thus get rid of the dissonance. Apparently, it is not what happens, since implicit and explicit attitudes are weakly correlated. This suggests that people don't experience dissonance in the first place, which is easy to explain if we assume that IAs are opaque. In this case, one is not aware of having inconsistent attitudes (cf. Berger 2020).

These considerations jointly support the claim that IAs are opaque. This, together with the belief view, entails that implicit beliefs are opaque. This sets them apart from ordinary beliefs. Nonetheless, the opaqueness of IAs shouldn't be exaggerated: it is not the case that one is *never* in a position to know them. For instance, one study has found people can quite accurately predict their IAT score prior to taking the test (Hahn et al. 2014). This indicates that self-knowledge of IAs is attainable in some circumstances[23]. That said, such attitudes are much more opaque than ordinary beliefs. So, the self-knowledge objection looms.

[23] Hanh et al. 2014 speculate that participants were able to detect their affective reactions to certain stimuli, which allowed them to predict their IAT score. If this interpretation is correct, this study at best shows that IAs are inferentially accessible. Accordingly, it doesn't constitute a counterexample to the claim that IAs are more opaque than ordinary beliefs. I am grateful to an anonymous referee for bringing this point to my attention.

Section 3: Are implicit beliefs unconscious?

In this section, I will consider a tempting way of dealing with the self-knowledge objection: to say that implicit beliefs are unconscious. I will point to some reasons to be skeptical about this response.

It is very natural to take "implicit" in "implicit attitudes" to mean unconscious: after all, this is how "implicit" was understood at the time when psychologists introduced the term "implicit social cognition" (Greenwald and Banaji 1995). However, this identification of the implicit with the unconscious has recently become controversial. To begin with, there is evidence that, in some circumstances, one can become aware of one's IAs (Hahn 2014, Nier 2005)[24]. Given that awareness is tightly linked to consciousness, this evidence is often taken to indicate that IAs are *not* unconscious (Gawronski et al. 2006).

Furthermore, the notion of conscious belief is itself problematic. For example, some worry that it makes little sense to talk about conscious beliefs: one might hold that a mental state is conscious just in case it has phenomenal character, and that beliefs don't have such character (Braddon-Mitchell and Jackson 2007). Alternatively, one can argue that consciously believing must be an event, and that believing can't be an event given its functional profile (Crane 2001). So, for those who deny the idea of conscious belief, saying that implicit beliefs are unconscious is not helpful, since, in their view, every belief is unconscious.

Of course, not everybody takes the notion of conscious belief to be problematic. For one thing, one can hold that beliefs have phenomenal character (Horgan and Tienson 2002). For another, some don't tie consciousness to phenomenal character – for example, higher-order accounts of

[24] As pointed out in footnote 16, this kind of evidence can be interpreted as only showing that IAs are inferentially accessible. And arguably, the fact that the agent infers from some cues that she is in a certain mental state doesn't necessarily mean that this state is conscious. So, one might think that the evidence in question doesn't refute the claim that IAs are unconscious states.

consciousness don't do so (Carruthers and Gennaro 2020). As an illustration, consider one version of the latter accounts: the higher-order thought approach to consciousness. Very roughly, it holds that a mental state is conscious if and only if it is the object of a higher-order thought and it causes this thought non-inferentially (Carruthers and Gennaro, 2020)[25]. Applying this approach to implicit beliefs, one can argue that, unlike ordinary beliefs, such beliefs don't give rise to higher-order thoughts. Accordingly, this renders implicit beliefs unconscious, which in turn accounts for their opaqueness (cf. Berger 2020). This move, however, makes one wonder why implicit beliefs differ from ordinary beliefs in this respect. One can't just say that implicit beliefs are not accompanied by higher-order thoughts and thus unconscious: it doesn't address the self-knowledge objection but just pushes the problem one step further. So, we need not only a developed notion of conscious belief, but also some explanation of why it is not applicable to implicit beliefs as opposed to their explicit counterparts[26].

For these reasons, substantial work needs to be done before one can address the self-knowledge objection by saying that implicit beliefs are unconscious. This doesn't mean that this project can't be carried out. However, given its difficulty, it might be reasonable to look for a different solution.

[25] The clause about non-inferential causation is needed to rule out cases in which one arrives at a higher-order thought via an indirect route, like a testimony of one's shrink or looking at one's brain scan. Prima facie, such states remain unconscious despite being accompanied by the relevant higher-order thoughts.

[26] The same problem arises for those who appeal to phenomenal character: they must explain why implicit beliefs don't have it.

Section 4: Believing that P without judging that P

In this section, I will present my account of implicit beliefs, which holds that implicit beliefs constitute a special kind of belief – beliefs that are not accompanied by the relevant judgments. In the next section, I will show how this account can address the self-knowledge objection.

There are at least two ways to understand judgement. Some take it to be a mental action that results in belief fixation. For example, Peacocke claims that "to make a judgment is the fundamental way to form a belief." (Peacocke 1998). In a similar vein, Crane defines judgment as the formation of belief (Crane 2001). Alternatively, "judgment" refers to a mental action that consists in entertaining a proposition and endorsing it. Another way to put it is to say that judgment is the mental analogue of assertion (Dummett 1973)[27]. This construal of judgment is broader than the first one, since it allows one to judge a proposition that one already believes.

I use "judgment" in the second sense. So, my account of the relationship between believing and judging can be formulated as follows: if one believes that P, then, when this belief is occurrent, one is apt to judge that P[28]. Now, it is clearly not always the case that occurrent beliefs are accompanied by the relevant judgments. When one buys groceries in the nearby supermarket, one's movements are guided by a multitude of beliefs about the locations of different products. But obviously, one needn't mentally assert these beliefs for them to have an impact on behavior. However, grocery beliefs can easily give rise to the relevant judgments: if one is asked whether pasta is stored next to cheese, one mentally asserts a certain proposition – the one that corresponds to one's belief about the whereabout of pasta. In other words, a belief

[27] It is tempting to think that judging so construed is simply an assertion without articulation: in other words, an assertoric utterance in inner speech. But this claim is problematic, for some have argued for the existence of thoughts not exhibiting any language-like qualities – so-called "unsymbolized thinking" (Hurlburt and Akhter 2008). If such thinking exists, the analogy between judgment and assertion is to be understood more loosely.

[28] Terminological note: Recall that by "occurrent" I mean beliefs that are operative, i.e., are carrying out their action- and thought-guiding role. One might want to define "occurrent believing that P" as judging that P. In this case, the presented conditional is a tautology. Alternatively, one can say that, if a belief that P is tokened and thus plays its action-guiding role, one judges that P even in the absence of the "inner assertion" that P (cf. Silins 2012). It should be clear that I have in mind different notions of judgment and occurrent belief.

can be operative without giving rise to the relevant judgment, but often the occurrent belief that P and judging that P go together[29].

By contrast, a tokening of an implicit belief doesn't give rise to the relevant judgment. This claim is stronger than simply saying that implicit beliefs often operate without causing the relevant judgments. The latter is true of grocery beliefs, but such beliefs are not implicit. Rather, the point is that the agent with the implicit belief P doesn't judge that P even when she is deliberating about P or related claims. Even if someone brings P to her attention, she won't mentally affirm this claim. To illustrate: Sam's implicit belief is being operative when he puts the dossier of R's member in the pile "Rejected". This action is partly caused by his belief that people of race R are less competent, P. However, Sam doesn't judge this to be so, for he sincerely endorses the claim that one's race is irrelevant for one's competence. If somebody asks him whether he thinks that people of race R are less competent, he will reject this claim. We can even stipulate that all the evidence available to Sam refutes P. In such circumstances, it is unlikely that he will judge that P.

One might wonder how P can affect Sam's mental states and actions if he never judges P. For one thing, implicit beliefs are shown to be responsible for micro-behaviors like eye contact or choosing seating distance (Dovidio et al. 1997, Wilson et al. 2000). Such behaviors are normally not caused by judgments: one doesn't, say, judges that such and such people are dangerous and, as a result, avoid looking at them. More likely, one's belief that these people are dangerous causes the relevant behavior bypassing judgment[30].

[29] Another possibility is judging without believing, in which case one mentally affirms something but fails to take it to heart and act accordingly (Cassam 2010). Such cases lend further support to the idea that believing and judging don't necessarily go together – they are separate mental states: judging is not necessary for believing, and believing is not necessary for judging.

[30] Levy 2015 argues that beliefs are not responsible for micro-behaviors, and for this reason, he denies that IAs are beliefs. See Mandelbaum 2016 for a reply.

In reply, one can say that Sam's decision not to hire a person of race R is not a micro-behavior, and that mental events like decisions tend to be preceded by judgments, like the judgment that a given candidate is less suitable for the job. How, the objection goes, can Sam make this judgment without judging that P at some point? This objection overintellectualizes decision-making. It tacitly assumes that this process works like an argument: one endorses certain propositions and, as a result of this endorsement, comes to accept another proposition, which is perceived to be supported by them. In this model, a belief can affect decision-making only if its content is endorsed, which amounts to saying that this belief causes the relevant judgment. This model ignores other ways in which a belief can impact decision-making. Here is one example of such an alternative route: a tokening of the belief that P consists in, among other things, a tokening of its constituent concepts, such as INCOMPETENT. This concept might be embedded in a network of associated concepts and valences, which would, in turn, get activated via established mechanisms of evaluative and semantic priming. Suppose that INCOMPETENT is linked to concepts like LAZY and is bound with negative affect. It is easy to see how activation of all these mental entities might bias Sam against a given candidate and thus result in the decision not to hire her.

One might wonder why it is the case that implicit beliefs are not accompanied by the relevant judgments. The answer, I contend, lies in the fact that implicit beliefs are evidentially problematic by one's own lights[31]. We can stipulate that Sam endorses egalitarian beliefs and

[31] One might worry that this consideration is not applicable to what Holroyd 2016 calls "harmony cases" and "protocol-adhering cases". Harmony cases are cases in which one's explicit and implicit attitudes align. Protocol-adhering cases are a subset of harmony cases. In such a scenario, a person with racist explicit and implicit attitudes sticks to protocol (when making a hiring decision, for example) and tries *not* to act on them. He succeeds only to some extent: his evaluations are guided by his intention not to discriminate, but his IA continues to affect his micro-behaviors without him being aware of that. One might argue that in cases like that, implicit beliefs are not evidentially problematic by one's own lights. Yet, in this scenario the implicit belief remains somewhat problematic by the agent's own lights in the following sense: even though the agent thinks that it is supported by evidence, he also thinks that this belief shouldn't be relied upon when hiring people. He might think that some considerations (prudential or having to do with fairness, say) speak against employing this belief in such situations. We can thus say that these considerations "defeat" the implicit belief in a given context. For this reason, the agent refrains from mentally affirming the content of this belief. But this of course doesn't prevent this belief from having an impact on the agent's micro-behaviors. As I will argue later, lack of relevant judgments explains the

repudiates racism because he has good reasons for doing so: maybe the evidence he possesses conclusively shows that one's race is irrelevant for one's professional qualification. For this reason, his implicit belief goes against his considered assessment of the subject matter[32]. We can say that Sam assents to the claim that one epistemically ought not believe that P. However, we stipulated that he believes P all the same. His belief is, therefore, akratic (Chislenko 2021). The akratic nature of the implicit belief P is, arguably, relevant to Sam's disposition to judge that P. A primary route to judgment is reasoning: one often judges that P because P is the outcome of reasoning on the agent's part. But patently, P can't be the outcome of Sam's reasoning because he takes his evidence to disconfirm P. For that reason, he is not in a position to engage in reasoning that terminates in P: on the contrary, his thinking about the subject matter in question leads him to the conclusion that not-P.

Reasoning is not the only route to judgment. One often judges that P as a result of having a perceptual experience that grounds P. But this option is not applicable to our case: there is no perceptual experience on which P is based. Another option is to say that judgment can be grounded in credal feelings that accompany entertaining a proposition. Simply put, P seems right to the agent, and that's why she judges that P – no reasoning is involved. However, credal feelings tend to be eliminated if one has access to defeaters for the proposition at hand. This is true of Sam, so he won't judge that P, since he has defeaters for this claim.

opacity of implicit beliefs. So, the fact that the agent doesn't endorse the content of his implicit belief in the protocol-adhering scenario, explains why he is not aware of this belief being operative. I am grateful to an anonymous referee for making me think about the application of my account of implicit beliefs to harmony and protocol-adhering cases.

[32] What if it doesn't? Arguably, in this case, one doesn't have an implicit belief – instead, one has an ordinary belief. Differently put, a person who endorses a racist claim is not an implicit racist – she is simply a racist.

Section 5: Judging as a route to one's beliefs

In the previous section, I argued that implicit beliefs have a peculiar feature, which distinguishes them from ordinary beliefs: they are not accompanied by the relevant judgments. In this section, I will show how this feature can explain the opaqueness of implicit beliefs.

The first thing to notice here is that one can often answer the question of whether she believes that P by answering the question whether P. This is the upshot of the following oft-cited passage from Evans 1982:

[I]n making a self-ascription of belief, one's eyes are, so to speak, or occasionally literally, directed outward — upon the world. If someone asks me 'Do you think there is going to be a third world war?,' I must attend, in answering him, to precisely the same outward phenomena as I would attend to if I were answering the question 'Will there be a third world war?' (Evans, 1982, 225)

Even though this observation must be qualified[33], many believe that it captures an important feature of self-ascriptions of belief, which is called "transparency": if one wants to know whether one believes that P, one must look not on the mental state in question (the belief that P), but through it (hence, the metaphor of transparency) – on the state of affairs that it represents, i.e., P.

One might agree that we often use the Evans-style procedure to self-ascribe beliefs. However, it is not immediately clear why this procedure results in beliefs enjoying a positive epistemic status. Several attempts have been made to explain this. Some of them ground positive epistemic status of self-beliefs in non-epistemic properties, such as one's agential authority

[33] To name a few caveats: Evans says that, to make correct self-ascriptions, one *must* attend to outward phenomena, but this is too strong – surely other routes, like inference from behavior, are available. Second, sometimes the Evans-style procedure issues the verdict "maybe P is the case", but one doesn't believe that P (think about the question of whether it will be raining in three weeks). For more on such caveats, see Silins 2012.

over one's beliefs (Moran 2001), whereas others point to epistemic dimensions of self-beliefs (Byrne 2005; Fernández 2003). I will focus on the latter variety of transparency accounts and discuss Byrne's account, but the same points could be made if I chose Fernandez's view. In a nutshell, I will argue that this kind of transparency accounts supports the claim that judgment is a route to one's beliefs. Differently put, one comes to know one's beliefs by attending to one's judgments[34].

Byrne claims that self-knowledge of belief is acquired by following, or trying to follow, this rule:

Bel: If P, believe that you believe that P.

To see how Bel works, let's consider the following example. John wants to find out whether he believes that Trump will be re-elected (T). If he were to follow Bel, he would examine the evidence that bears on the truth of T. Suppose that this evidence supports T. In this case, John would judge that T and thus come to form a certain second-order belief: the belief that he believes that T. Byrne thinks that Bel accounts for two features of self-knowledge of beliefs. First, such self-knowledge is peculiar: it is obtained by the method(s) that differ from the method(s) one uses to obtain knowledge of other people's beliefs. Second, self-knowledge of beliefs is privileged, since the method(s) of obtaining it are more reliable than the methods by which one comes to know other people's beliefs. Clearly, Bel can account for the peculiarity of self-knowledge, since Bel is applicable only to oneself: if I take it to be the case that T, I shouldn't conclude that another person believes that T, since she might simply not have access to the considerations on which I have based my judgment about T[35].

[34] Fernández 2003 claims that the belief that P and the belief that one believes that P have the same basis: for example, the same perceptual experiences as of P grounds both the first-order belief and the second-order belief. So, his account supports my claim that one knows one's beliefs by consulting one's judgments, since one's judgments are sensitive to the factors that ground one's first-order beliefs.

[35] Sometimes we can use this procedure to figure out what other people believe. Suppose I come to your house and notice that the walls are freshly painted. In this case, I have justification to conclude that you believe that the

Bel also explains why self-knowledge of beliefs is privileged. To apply Bel, one has to figure out whether the condition, P, that is specified in its antecedent obtains. To do so, one has to make a judgment about P. But making such a judgment amounts to taking this proposition to be true, which amounts to believing it. Suppose one judges that P. In this case, Bel requires that she believes that she believes P. This second-order belief turns out to be correct because the agent believes that P: if she didn't, she wouldn't have judged that P and thus wouldn't have applied the rule in the way she did. Therefore, the very application of Bel ensures that the rule outputs true second-order beliefs.

Now, if Bel explains why self-knowledge is peculiar and privileged, it is tempting to conclude that self-knowledge is indeed acquired in the way that is specified by the rule. If so, beliefs are transparent[36]. This has an important consequence for our topic of the opaqueness of implicit beliefs: such beliefs are not transparent, i.e., not generated by transparency methods, and this sets them apart from explicit beliefs. And I have already mentioned the reason why implicit beliefs are not transparent: they depart from one's judgments about the relevant propositions. So, if one uses one's judgments to come to know one's first-order beliefs, as Bel requires to do, one is bound to self-ascribe incorrect beliefs[37]. In other words, the fact that implicit beliefs are not transparent explains – pardon the pun – why they are opaque.

---

walls are freshly painted. However, to make this inference, I need some background information: that you live in this house, that it is unlikely that the walls were painted without your permission, etc. By contrast, I don't need all this information to make the correct self-ascription (Silins 2012).

[36] Bel and other transparency accounts are open to pressing objections. For one thing, Bel doesn't work with irrational beliefs, like the belief that spilling salt brings bad luck (Gertler 2011): one can believe it despite not endorsing this claim. However, given that beliefs tend to be evidence-responsive, irrational beliefs are the exception rather than the rule. For this reason, beliefs *tend* to align with judgments, and this makes Bel a useful – albeit not infallible – route to self-knowledge. For another, it has been argued that proper application of Bel presupposes self-knowledge of some mental states. It means that Bel is not fully transparent rule (Gertler 2011). This objection doesn't threaten the argument in this section because this argument doesn't hinge on the claim that Bel is fully transparent.

[37] For the present purposes, it doesn't matter whether one infers that one believes that P from one's judging that P, or whether this self-knowledge counts as immediate. What matters is that the link between judging and believing puts one into a good epistemic position with respect to one's beliefs. Second, this route is available only if we have reasons to think that one is in a good epistemic position with respect to one's judgments. This claim doesn't seem to be very controversial, for judgments, arguably, have certain phenomenology, which grounds their self-knowledge. One reason to think that judgments have phenomenology comes from the fact that judgments are

Let me recap the dialectic so far. I have argued that implicit beliefs are not accompanied by the relevant judgments. This feature explains their opaqueness because judgment is an important route to one's beliefs. To support the latter claim, I discussed the idea that beliefs are transparent as well as some explanations of why self-beliefs generated by transparency methods enjoy a positive epistemic status. In particular, I mentioned Byrne's account, which features the rule Bel. The discussion of Bel highlighted that this rule works because belief is tightly linked to judgment. The fact that implicit beliefs are not accompanied by the relevant judgments means that these beliefs are not transparent: one can't know them by means of Bel or some other method accommodating transparency. This explains one's poor epistemic position with respect to them[38].

Section 6: Belief, judgment, evidence-responsiveness

One can raise the following objection to the paper's argumentative strategy: The paper argues that implicit beliefs are not accompanied by the relevant judgments, and that this fact explains the opaqueness of implicit beliefs because judgment is a route to one's belief. But if the link

mental events, like paradigm examples of states with phenomenology, such as sensory and pain experiences. Second, it is hard to make sense of the idea of unconscious judgment. One explanation for this is phenomenology: judging that P is accompanied by a certain phenomenology, and the presence of this phenomenology explains why one knows that one judges P. One might then appeal to the phenomenology of judgment to ground self-knowledge of belief. This appeal is not uncontroversial because judging is an instance of thinking, as opposed to perception, and some people are skeptical about cognitive phenomenology (roughly, the phenomenology of thinking). But one can invoke the phenomenology of judgement without committing oneself to the idea of distinctive cognitive phenomenology. Instead, one can hold that the phenomenology of judgment comes from sensory or quasi-sensory states that accompany acts of judging (Carruthers 2005). One can also argue that inner speech (and therefore judgment) tends to be associated with agentive phenomenology, whatever exactly this phenomenology amounts to – the feeling of control, the feeling of mental effort or something else. This distinctive agentive phenomenology might be part of what grounds self-knowledge of judgment.

[38] This reasoning doesn't conclusively show that implicit beliefs are opaque. One can argue that such beliefs can be known via inference from behavior and a variety of mental states that are caused by a propositional attitude (Lawlor 2009). However, this route might not work when it comes to implicit beliefs. Such beliefs tend to be at odds with one's self-image, and for this reason, one might be motivated *not* to discover that one has them. So, this route becomes subject to motivated reasoning (Kunda 1990). Similarly, implicit beliefs are an obvious target of a psychological immune system (Gilbert et al. 1998). Pursuing this argument is a project for a different paper.

between judging and believing is severed for implicit beliefs, why is it preserved for ordinary beliefs? Clearly, I can't say that judging is a necessary feature of believing. To meet this objection, I must provide an account of belief that, on the one hand, explains why beliefs tend to be accompanied by judgments and, on the other hand, leaves room for beliefs that are *not* accompanied by judgments[39].

Moreover, on my account, implicit beliefs fail to respond to evidence. This feature is crucial for explaining why they depart from judgments, for judgments, arguably, tend to be more evidence-responsive than beliefs[40]. So, if beliefs and judgments are to be linked, beliefs must align with evidence. For this reason, by giving a story about why certain beliefs are not aligned with evidence, one *ipso facto* explains why certain beliefs lack a connection with judgments.

We can appeal to teleology to explain belief's evidence-responsiveness (and thus its link to judgment). One can argue that belief's function is to help an organism to select actions conducive to the satisfaction of its desires. To fulfil this function, beliefs must carry information

[39] I am grateful to an anonymous referee for pressing me to articulate an account of belief that is compatible with the main argument of the paper.

[40] One might wonder why judgments are more evidence-responsive than beliefs. Here are some considerations supporting this claim. To begin with, the central case discussed in the paper suggests it is true: in accordance with available evidence, Sam judges that race R is irrelevant for one's competence, but he believes that job candidates of race R are less competent than others. This scenario indicates that beliefs are less likely to rationally respond to evidence than judgments are. Sam's case is not an isolated example of this pattern: some empirical studies show that presenting new evidence affects participants' deliberate judgments about the target object, whereas their IAs remain virtually the same (Wilson et al. 2000). If one accepts that IAs are beliefs, this evidence gives reason to think that beliefs are more sluggish than judgments when it comes to the uptake of evidence. This sluggishness of belief makes sense in light of the idea that belief storage is fragmented (see below). Fragmented storage is relevant for belief update: some accounts of fragmentation hold that a belief gets updated only if the fragment hosting this belief is active (Bendana and Mandelbaum 2021). So, we might have the following situation: the agent who believes that P encounters strong evidence against P; however, the fragment hosting her belief that P happens not to be open, so this belief fails to get updated. By contrast, her judgment is not affected by fragmentation in this way. When she sees good evidence against P, she is likely to mentally affirm not-P. This doesn't mean that people always judge in accordance with evidence. It is easy to imagine a person who manages to ignore (or re-interpret) available evidence for P and thereby fails to judge that P. There are several mechanisms that can explain this bias in judgment – for example, motivated reasoning and a psychological immune system (Kunda 1990, Gilbert et al. 1998). But the important point is this: these mechanisms have an impact on how one handles evidence, so they also have an impact on belief formation and update. And in addition, belief formation and update are affected by fragmentation. This explains why beliefs are less evidence-responsive than judgments. I thank an anonymous referee for pushing me to provide some support for this latter claim.

about the environment, or track reality (Papineau 2013). Arguably, evidence is the best guide to how things are, so reality tracking explains evidence-responsiveness.

But from this, it doesn't follow that beliefs *always* respond to evidence. Belief fixation and belief maintenance are inevitably subject to constraints stemming from our cognitive architecture. One such (putative) constraint is the ballistic nature of belief fixation: a tokening of any truth-apt mental representation P inevitably results in believing P (Gilbert 1990, Mandelbaum 2014). On this view, a person might acquire lots of unfounded beliefs if she doesn't reject them right away. Such a scenario is possible because rejection is argued to be an effortful cognitive operation that is easily disrupted.

Another constraint follows from fragmentation of belief: roughly, the idea that beliefs are stored, as it were, in different folders, or fragments (Bendana and Mandelbaum 2021) [41]. A belief gets retrieved and poised to guide behavior once the fragment hosting it is activated. Crucially for our purposes, the same is true of belief update: a belief gets updated when the hosting fragment is activated. Fragmentation can thus explain why a belief is maintained in the face of counterevidence. For one thing, tokens of the same belief type can be hosted in many fragments. Accordingly, encountering counterevidence might eliminate only some of these tokens. For another, one can speculate that some fragments get activated in special circumstances, which means that their contents get updated only in those situations.

Overall, these considerations explain the link between belief and evidence-responsiveness. At the same time, they account for the recalcitrance of certain beliefs. It in turn explains why, even though beliefs tend to be accompanied by judgments, some beliefs fail to do so. All this

[41] The talk about storage implies representationalism about belief, but some versions of fragmentation don't presuppose it. For example, one can hold that the dispositions that are constitutive of a particular belief are indexed by certain circumstance C – that is, apt to become manifested in C (Elga and Rayo 2021).

suggests a direction in which one can go to build an account of belief compatible with the claims about implicit beliefs that I defend.

Section 7: Further advantages of the present account of implicit beliefs

Up until now, I have argued that my account of implicit beliefs as beliefs that are not accompanied by the relevant judgments explains why such beliefs are more opaque than ordinary beliefs. In this section, I will discuss other advantages of this account.

First, my proposal can account for the differences between the functional profile of explicit and implicit beliefs. It has been pointed out that explicit and implicit attitudes are predictive of different types of behavior: EAs guide controlled responses, such as verbal behavior or reasoning, whereas IAs contribute to more automatic behaviors like body language (Dovidio et al. 1997, Wilson et al. 2000). If one thinks that EAs and IAs are beliefs, one has to explain why states of the same type have these distinct functional profiles. My proposal has resources to account for this difference. Since implicit beliefs are not accompanied by the relevant judgments, such beliefs can't contribute to one's cognitive economy in *the same way* in which their explicit counterparts do. Implicitly believing that P doesn't give rise to the judgment that P, and for this reason, the agent is not expected to assert P or to use P as premise in her reasoning. This explains why implicit beliefs don't feature in one's reasoning and speech. On the other hand, the absence of judgment doesn't prevent this implicit belief from affecting one's cognition and behavior in some other ways because beliefs don't always affect behavior via judgments.

Second, the account in question might explain why implicit beliefs are isolated from the rest of the agent's beliefs. To see this, we need to go back to the idea of fragmentation of belief introduced earlier. As an illustration, consider a person who believes three inconsistent

propositions: that Nassau Street runs east-west, that the railroad nearby runs north-south, and that these street and railroad are parallel (Lewis 1982). Lewis suggests that the agent's belief system is broken into fragments, such that different fragments come into action in different situations. No fragment contains the conjunction of the three propositions, which explains why this inconsistency remains unnoticed up to a certain point.

This framework has been applied to implicit beliefs (Bendana and Mandelbaum 2021). It has been argued that such beliefs and their explicit counterparts are stored in different fragments, which explains why agents act on different beliefs in different situations. However, as Lewis's example suggests, these fragments are not completely cordoned off from each other: after all, Lewis realized that he held inconsistent beliefs and eliminated this inconsistency. By contrast, the inconsistencies between implicit and explicit beliefs persist. This suggests that the fragments in which implicit beliefs are stored are isolated from the fragments with explicit beliefs. My account can explain why it is so. Implicit beliefs bypass judgment, so the agent who implicitly believes that P and explicitly believes that not-P is not in a position to judge P and not-P. For this reason, she is not in a position to notice this inconsistency and do something about it. Arguably, this is how Lewis adjusted his belief system: he noticed that he endorsed the three inconsistent claims, and this prompted him to revise his beliefs. This route to belief revision is not available for the agent with implicit beliefs[42].

Conclusion

There are good reasons to think that IAs – mental states responsible for implicit bias – are ordinary beliefs. This view faces an objection stemming from the fact that IAs don't seem to

[42] It doesn't mean that becoming aware of an inconsistency suffices to eliminate it. For it might turn out that implicit beliefs with the same content are stored in more than one fragment. So, eliminating one such belief doesn't necessarily make one's attitudes aligned.

behave like ordinary beliefs do. One such difference in behavior has to do with self-knowledge: one tends to be in a good epistemic position with respect to one's beliefs; by contrast, one's epistemic position vis-à-vis IAs is poor. How can it be the case if IAs are beliefs? This paper addresses this worry and thereby provides indirect support for the belief view. I argued that implicit beliefs constitute a special kind of belief – beliefs that are not accompanied by the relevant judgments. This feature explains one's poor epistemic position vis-à-vis implicit beliefs because judgment is an important route to self-knowledge. This latter claim is supported by the transparency of belief. In addition to dealing with what I called "the self-knowledge objection" to the belief view, the paper discusses several advantages of this account of implicit beliefs.

Acknowledgments

For helpful feedback on previous drafts of this paper, I would like to thank the audiences at the University of Barcelona, LOGOS, and the University of Antwerp, Centre for Philosophical Psychology, as well as two anonymous referees. I am also grateful to participants of the 30th European Society for Philosophy and Psychology Conference, of the Graduate Reading Group at the University of Barcelona and of the Taller d'Investigació en Filosofia at the University of Valencia. Finally, I am greatly indebted to Pepa Toribio for her invaluable help and insightful comments.

Funding

Research for this article was supported by MICIU/AEI/ 10.13039/501100011033 and by ERDF A way of making Europe, under grant agreement PID2021-124100NB-I00,

by AGAUR, under grant agreement 2021-SGR-00276, and by Grant CEX2021-001169-M funded by MICIU/AEI/10.13039/501100011033.

References

Bendana, J. & Mandelbaum, E. (2021) *The fragmentation of belief,* in Borgoni, C., Kindermann, D., & Onofri, A. (eds.) *The fragmented mind*, Oxford University Press, pp. 78-108.

Berger, J. (2020) Implicit attitudes and awareness. *Synthese*, *197*(3), pp. 1291–1312. https://doi.org/10.1007/s11229-018-1754-3.

Boucher, K. L., & Rydell, R. J. (2012) Impact of negation salience and cognitive resources on negation during attitude formation. *Personality and Social Psychology Bulletin, 38*(10), pp. 1329–1342.

Braddon-Mitchell, D. & Jackson, F. (2006) *Philosophy of mind and cognition: An introduction*, Blackwell.

Briñol, P., Petty, R., & McCaslin, M. (2009) Changing Attitudes on Implicit versus Explicit Measures: What is the Difference? In R. Petty, R. Fazio, and P. Brinol (eds.) *Attitudes: Insights from the New Implicit Measures*, New York: Psychology Press

Brownstein, M. (2018) *The Implicit Mind: Cognitive Architecture, the Self, and Ethics*, New York: Oxford University Press.

Byrne, Alex (2005) 'Introspection'. *Philosophical Topics* 33, pp. 79-104.

Cassam, Q. (2010) Judging, believing and thinking. *Philosophical Issues*, *20*(1), pp. 80–95. https://doi.org/10.1111/j.1533-6077.2010.00179.x

Carruthers, P. (2005) Conscious experience versus conscious thought, in *Consciousness: Essays from a Higher-Order Perspective*, Oxford: Oxford University Press, pp. 134–57.

Carruthers, P. (2009) An architecture for dual reasoning, in: Evans JSBT, Frankish K. (eds.) *In two minds: dual processes and beyond,* Oxford University Press, Oxford, pp. 109–127.

Carruthers, P. & Gennaro, R. (2020) *Higher-order theories of consciousness*, [Online], Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/entries/consciousness-higher/ [10 September 2024]

Chislenko, E. (2021) How can belief be akratic? *Synthese*, *199*(5–6), pp. 13925–13948. https://doi.org/10.1007/s11229-021-03404-0.

Cone, J., Mann, T. C., & Ferguson, M. J. (2017) Changing our implicit minds: How, when, and why implicit evaluations can be rapidly revised, *Advances in Experimental Social Psychology*, pp. 131–199. https://doi.org/10.1016/bs.aesp.2017.03.001

Cooper, J. (2007) *Cognitive dissonance: Fifty Years of a classic theory*, Sage Publications.

Crane, T. (2001) *Elements of mind: An introduction to the philosophy of mind*, Oxford University Press.

De Houwer, J. (2014) A propositional model of implicit evaluation, *Social and Personality Psychology Compass*, *8*(7), pp. 342–353. https://doi.org/10.1111/spc3.12111

Dovidio, J. F., Kawakami, K., Johnson, C., Johnson, B., & Howard, A. (1997) On the nature of prejudice: Automatic and controlled processes, *Journal of Experimental Social Psychology* 33: pp. 510–540.

Dummett, M. (1973) *Frege: Philosophy of Language*, London: Duckworth

Elga, A. & Rayo., A. (2021) Fragmentation and information access, in Borgoni, C., Kindermann, D., & Onofri, A. (eds.). *The fragmented mind,* Oxford University Press, pp. 37-54.

Evans, G. (1982) *The Varieties of Reference*, Oxford University Press.

Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986) On the automatic activation of attitudes. *Journal of Personality and Social Psychology*, *50*(2), pp. 229–238. https://doi.org/10.1037/0022-3514.50.2.229.

Fernández, J. (2003) Privileged Access Naturalized, *The Philosophical Quarterly*, 53 (2003), pp. 352-72.

Frankish, K. (2016) Playing Double: Implicit Bias, Dual Levels, and Self-Control, in Brownstein, M. & Saul, J. (eds.) *Implicit Bias and Philosophy, Vol. 1: Metaphysics and Epistemology*, Oxford: Oxford University Press, pp. 23–46.

Gawronski, B. & Bodenhausen, G.V. (2006) Associative and Propositional Processes in Evaluation: An Integrative Review of Implicit and Explicit Attitude Change, *Psychological Bulletin*, 132(5), pp. 692–731.

Gawronski, B., Deutsch, R., Mbirkou, S., Seibt, B., & Strack, F. (2008) When "just say no" is not enough: Affirmation versus negation training and the reduction of automatic stereotype activation, *Journal of Experimental Social Psychology, 44*(2), pp. 370–377.

Gawronski, B., Hofmann, W., & Wilbur, C. J. (2006) Are “implicit” attitudes unconscious? *Consciousness and Cognition*, 15(3), pp. 485–499.

Gendler, T. S. (2008) Alief and belief, *Journal of Philosophy* 105(10), pp. 634-663.

Gertler, B. (2011) Self-Knowledge and the Transparency of Belief, in: A. Hatzimoysis (ed.) *Self-Knowledge*, Oxford: Oxford University Press, pp. 125–145

Gilbert, D. (1991) How mental systems believe. American Psychologist, 46(2), 107–119. https://doi.org/10.1037//0003-066x.46.2.107

Gilbert, D. T., Pinel, E. C., Wilson, T. D., Blumberg, S. J., & Wheatley, T. P. (1998) Immune neglect: A source of durability bias in affective forecasting, *Journal of Personality and Social Psychology*, *75*(3), pp. 617–638. https://doi.org/10.1037//0022-3514.75.3.617

Gregg, A., Seibt, B. & Banaji, M. (2006) Easier Done than Undone: Asymmetry in The Malleability of Implicit Preferences, *Journal of Personality and Social Psychology* 90 (1), pp. 1-20

Greenwald, A. G. & Banaji, M. R. (1995) Implicit social cognition: Attitudes, self-esteem, and stereotypes, *Psychological Review*, 102(1), pp. 4–27.

Greenwald, A., M. Banaji, & B. Nosek. (2015) Statistically small effects of the implicit association test can have societally large effects, *Journal of Personality and Social Psychology* 108, pp. 553–561.

Greenwald, A. G. & Lai, C. K. (2020) Implicit social cognition, *Annual Review of Psychology*, *71*(1), pp. 419–445. https://doi.org/10.1146/annurev-psych-010419-050837.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998) Measuring individual differences in implicit cognition: The Implicit Association Test, *Journal of Personality and Social Psychology*, *74*(6), pp. 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464.

Hahn, A., Judd, C. M., Hirsh, H. K., & Blair, I. V. (2014) Awareness of implicit attitudes, *Journal of Experimental Psychology: General*, 143(3), pp. 1369–1392. https://doi.org/10.1037/a003502.

Holroyd, J. (2016). VIII—what do we want from a model of implicit cognition? *Proceedings of the Aristotelian Society*, *116*(2), pp. 153–179. https://doi.org/10.1093/arisoc/aow005

Holroyd, J. & Sweetman, J. (2016) The heterogeneity of implicit bias. *Implicit Bias and Philosophy, Volume 1*, pp. 80–103. https://doi.org/10.1093/acprof:oso/9780198713241.003.0004

Horgan, T. & Tienson, J. (2002) The Intentionality of Phenomenology and the Phenomenology of Intentionality, in Chalmers, D. (ed.), *Philosophy of Mind: Classical and Contemporary Readings*, Oxford UP.

Hurlburt, R. T. & Akhter, S. A. (2008) Unsymbolized thinking. *Consciousness and Cognition: An International Journal, 17*(4), pp. 1364–1374. https://doi.org/10.1016/j.concog.2008.03.021

Johnson, G.M. (2020) The Structure of Bias, *Mind* 129/516, pp. 1193–236.

Johnson, I. R., Kopp, B. M., & Petty, R. E. (2016) Just say no! (and mean it): Meaningful negation as a tool to modify automatic racial attitudes, *Group Processes & Intergroup Relations*, *21*(1), pp. 88–110. https://doi.org/10.1177/1368430216647189

Karlan, B. (2021) The rational dynamics of implicit thought, *Australasian Journal of Philosophy*, *100*(4), pp. 774–788. https://doi.org/10.1080/00048402.2021.1936581

Kunda, Z. (1990) The case for motivated reasoning, *Psychological Bulletin*, *108*(3), pp. 480–498. https://doi.org/10.1037/0033-2909.108.3.480.

Kurdi, B. & Dunham, Y. (2020) Propositional accounts of implicit evaluation: Taking stock and looking ahead. Social Cognition, 38 (Suppl), pp. 42-67. https://doi.org/10.1521/soco.2020.38.supp.s42

Kurdi, B., & Dunham, Y. (2021). Sensitivity of implicit evaluations to accurate and erroneous propositional inferences. *Cognition*, 214, Article 104792. https://doi.org/10.1016/j.cognition.2021.104792

Lawlor, K. (2009) Knowing what one wants, *Philosophy and Phenomenological Research*, *79*, pp. 47–75. https://doi.org/10.1111/j.1933-1592.2009.00266.x

Levy, N. (2015) Neither fish nor fowl: Implicit attitudes as patchy endorsements, *Noûs*, 49(4), pp. 800–823. https://doi.org/10.1111/nous.12074

Lewis, D. (1982). Logic for Equivocators. *Noûs*, 16 (3), pp. 431–441.

Machery, E. (2016) De-Freuding Implicit Attitudes, in Brownstein, M., & Saul J. (eds.) *Implicit bias and Philosophy*, vol. 1, pp. 104-129

Madva, A. (2016) Why implicit attitudes are (probably) not beliefs, *Synthese*, *193*(8), pp. 2659–2684. https://doi.org/10.1007/s11229-015-0874-2

Mandelbaum, E. (2013) Thinking is believing, *Inquiry*, *57*(1), 55–96. https://doi.org/10.1080/0020174x.2014.858417

Mandelbaum, E. (2016) Attitude, association, and inference: On the propositional structure of implicit bias, *Noûs* 50(3), pp. 629-658

Mann, T. C., Kurdi, B., & Banaji, M. R. (2020) How effectively can implicit evaluations be updated? using evaluative statements after aversive repeated evaluative pairings. *Journal of Experimental Psychology: General*, *149*(6), pp. 1169–1192. https://doi.org/10.1037/xge0000701.

Monteith, M. J., Voils, C. I., & Ashburn-Nardo, L. (2001) Taking a look underground: Detecting, interpreting and reacting to implicit racial biases, *Social Cognition* 19(4), pp. 395–417.

Moran, R. (2001) *Authority and Estrangement: an essay on self-knowledge*. Princeton University Press.

Nanay, B. (2021) Implicit bias as mental imagery, *Journal of the American Philosophical Association*, *7*(3), pp. 329–347. https://doi.org/10.1017/apa.2020.29.

Nier, J. (2005) How dissociated are implicit and explicit racial attitudes? A bogus pipeline approach, *Group Processes & Intergroup Relations*, 8, pp. 39–52.

Nosek BA, Banaji MR. (2001) The go/no-go association task, Social Cognition 19(6), pp. 625–66.

Oswald, F., Mitchell G., Blanton, H., Jaccard J., & Tetlock, P. (2013) Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies, *Journal of Personality and Social Psychology* 105, pp. 171– 192.

Papineau, D. (2013) *There are no norms of belief,* in T. Chan (ed.) *The Aim of Belief*, Oxford University Press, pp. 64–79.

Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, B. D. (2005) An inkblot for attitudes: Affect misattribution as implicit measurement, *Journal of Personality and Social Psychology*, *89*(3), pp. 277–293. https://doi.org/10.1037/0022-3514.89.3.277.

Schwitzgebel, E. (2023) *Belief*, [Online], Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/entries/belief/ [10 September 2024]

Silins, N. (2012) *Judgement as a guide to belief,* in Smithies, D., & Stoljar, D. (eds.) *Introspection and consciousness*. Oxford University Press.

Stich, S. P. (1978) Beliefs and subdoxastic states, *Philosophy of Science*, *45*(4), pp. 499–518. https://doi.org/10.1086/288832

Sullivan-Bissett, E. (2019) Biased by Our Imaginings, *Mind and Language* 34/5, pp. 627–47.

Toribio, J. (2018) Implicit bias: From social structure to representational format. *THEORIA. An International Journal for Theory, History and Foundations of Science*, *33*(1), pp. 41-60. https://doi.org/10.1387/theoria.17751

Wilson, T. D., Lindsey, S., & Schooler, T. Y. (2000) A model of dual attitudes, *Psychological Review*, *107*(1), pp. 101–126. https://doi.org/10.1037/0033-295x.107.1.101.

Chapter 3: Associative inferential transitions, or One problem with Siegel's Response Hypothesis

Article published in *Acta Analytica* (2025), Online First

Abstract

What is inference? This paper discusses a recent account that aims to answer this question – Susanna Siegel's Response Hypothesis. The hallmark of inference, on Siegel's account, is the epistemic dependence of a mental transition's output state(s) on its input state(s). In this paper, I argue that some alleged non-inferential transitions exhibit the kind of epistemic dependence that Siegel's account takes to be characteristic of inference. More precisely, I argue that some associative transitions exhibit this kind of epistemic dependence—a problematic conclusion, since Siegel takes inference and association to be mutually exclusive kinds of mental transitions. I then suggest a way out of this problem: to reject the assumption that association and inference are mutually exclusive. This may be considered a bold move, for associative transitions are often considered a paradigm example of non-inferential transitions. So, I end up discussing the motivation behind the move and arguing that it opens up an attractive niche for the development of some philosophical projects.

Keywords: inference, association, epistemic justification, externalism, internalism

Introduction

Transitions between mental states come in many varieties. Smelling an apple pie can make us think of our grandmother just because our grandma often prepared this dish when we visited her in the past and not because there is any logical connection between apple pies and grandmothers. In other cases, transitions between mental states are more like arguments: the thought that John left the house without his keys follows from seeing John's key on the desk after he left. These are just two examples of how thinking might proceed. The examples are familiar, and it is easy to find the words to describe them: the former is an associative transition, whereas the latter is an instance of inferring. Although people seem to have an intuitive grasp on the notions, we need a more precise answer to the question of what distinguishes inference from association.

The main reason why a more precise answer is needed is that the notions of inference and associative transition play an important theoretical role both in philosophy and cognitive science. For example, they are brought up in the debate about the structure of the mind. Some philosophers and cognitive scientists argue that the processes involved in thinking belong to two different kinds. One kind, System 1 processes, is often equated with associative processes, whereas the other kind, System 2 processes, overlaps with inference to a large extent (Frankish & Evans 2009, Kahneman 2011).

Or take the important role that the notion of inference plays in epistemology, where it is closely linked to the idea of the basing relation and epistemic responsibility (Boghossian 2018). Since, it is often argued, one belief is based on some other belief just in case the former is inferred from the latter, establishing the basis of a belief is crucial for assessing its justificatory status – and, clearly, the notion of justification is central to epistemology. As far as epistemic responsibility is concerned, one can argue that a person can be held responsible for having

reasoned well or badly. The notion of reasoning seems to be very close to the notion of inference, so clarifying the latter is relevant for the responsibility issue.

Given these important theoretical roles, it should be clear why one should want to offer a precise account of inference. The aim of the paper is to examine in detail the account recently proposed by Susanna Siegel, which she calls the Response Hypothesis (Siegel, 2017, Siegel, 2019). In a nutshell, the Response Hypothesis (RH henceforth) amounts to the following claim: a transition between two mental states is inferential if and only if the input state causes the output state, and the output epistemically depends on the input. I will argue that RH faces what we might call "the demarcation problem": it fails to demarcate inferential from non-inferential transitions, associative transitions being their prime example. The problem arises because associative transitions can also exhibit epistemic dependence characteristic of inference. Given that she takes inference and association to be mutually exclusive, associative transitions that exhibit epistemic dependence are a counterexample to her account.

The plan for the paper is as follows. In Section 1, I give some examples of inferential and associative transitions and provide an account of association. In Section 2, I introduce RH and spell out the notion of epistemic dependence. Here I also look at some advantages of RH and briefly compare it with some competing accounts. In Section 3, I discuss associative transitions exhibiting epistemic dependence and thus constituting counterexamples to RH. In Section 4, I discuss some objections to these counterexamples. Finally, in Section 5, I offer one way in which RH can accommodate these counterexamples: to bite the bullet and accept that some associative transitions can be inferential. This may be considered a counterintuitive move, so I discuss the motivation behind it and argue that it offers an attractive niche for some philosophical projects.

Section 1: **Inferential and associative transitions: brief introduction**

Let's begin with a clear example of inference. A detective investigates a murder scene. She is told by the coroner that, given the location of the wound on the victim's body, the murderer must have been quite short, one meter sixty or so. Smith is the only suspect who is so short. The detective reasons as follows:

(Whodunnit) The murderer is one meter sixty tall. Of all the suspects, only Smith is one meter sixty tall. Smith is the murderer.

This sequence of thoughts is an example of an inferential transition. Here the detective adopts a new belief based on her other beliefs. I take it that chains of thoughts like Whodunnit are paradigm examples of what we pre-theoretically call 'inference'[43].

What about association? To a first approximation, a transition between two concepts, or thoughts featuring these concepts, is associative just in case the transition is driven by the concepts' co-occurrences in the past. Suppose that thinking about one's grandmother makes one think about an apple pie. If one moves from a grandmother-thought to a pie-thought because the concepts GRANDMOTHER and APPLE PIE were repeatedly co-activated in one's head, then the transition is associative.

Inferential and associative transitions don't exhaust the space of mental transitions, but I will focus on just these two. I do so for two reasons. First, lots of debates in cognitive science and philosophy revolve around them – the reflection of the assumption that inference and association constitute a significant chunk of mental processes. To see this, consider a distinction between System 1 and System 2 processes. The former are fast, automatic, not cognitively demanding and associative, whereas the latter are the opposite – slow, controlled, cognitively demanding and rule-based (Kahneman 2011). Now, it would be an oversimplification to equate

[43] Inferring does not necessarily involve beliefs. For example, one can accept a claim for the sake of argument and look at what follows from it. An inference of this kind doesn't start with a belief. That said, the discussion will revolve around inferences that feature beliefs.

System 2 processes with inferential processes. However, these two notions largely overlap. For example, the deliberate reasoning on the part of the detective in Whodunnit is a System 2 process. This overlap and the identification of System 1 processes with associative processes (Kahneman 2011) show that inference and association play an important role in theorizing about the mind. This makes their explication worthwhile.

Second, among the prime examples of non-inferential transitions we encounter many associative transitions. Moreover, many discussions of inference assume that inference and association are mutually exclusive: if a process is inferential, it is not associative, and the other way around (Boghossian 2014, Quilty-Dunn & Mandelbaum 2017, Siegel 2017, Siegel 2019). Whether one accepts this claim or just believes that many instances of association are clear examples of non-inferential transitions, each position makes association a perfect test for an account of inference: if such an account can't demarcate inference from association, the account is in trouble and needs to be modified.

So far, I have mainly clarified the notions of inference and association by giving examples. It is much harder to provide their strict definitions. When introducing the notion of association, I mentioned an intuitive idea that co-occurrences are a necessary and sufficient condition of forming an association. However, it has been found that such co-occurrences are neither necessary nor sufficient (Rescorla 1988). Consider the phenomenon of blocking (Kamin 1969): if stimulus A is associatively linked with stimulus X, pairing a stimulus made up of A and B with a stimulus X won't result in a link between B and X. The existing link between A and X blocks the formation of a new link; hence, the term "blocking". Blocking thus demonstrates that co-occurrences are not sufficient for forming an association.

In addition to blocking, there are other findings that complicate the task of defining association[44]. So, for the purpose of the present discussion, I will provide a stipulative definition that is applicable to a sizeable number of associative transitions: a transition from representation A to representation B counts as association just in case the stimuli represented by A and B co-occurred in one's experience and, as a result of such co-occurrences, a tokening of one representation causes a tokening of the other[45].

Providing an account of inference is the objective of the next section. However, at this point I can mention a feature of inference that is taken to be necessary by many accounts: the causal connection between the premise(s) and the conclusion(s)[46]. It is clear, however, that causation alone can't demarcate inference from association because in an associative transition, the input state also causes the output state. The question is then what other features of inference set it apart from non-inferential transitions.

Section 2: Siegel's Response Hypothesis

There is no consensus about the definition of inference. Some argue that inferring is a mental action in which the subject takes the premises to support the conclusion and draws this conclusion because of that fact (Boghossian 2014). A natural construal of the taking state holds that one is aware of it or can become aware of it (Boghossian 2014). Others deny that such

[44] These include overshadowing, second-order conditioning and sensory pre-conditioning (Bouton 2016).

[45] The word "associative" is used in at least two different ways. It can refer to links between representations that are formed as a result of their co-activation. This is the way I use the word throughout the paper. On the other hand, one might encounter claims to the effect that all cognitive processes are implemented by associative networks made up of nodes connected to each other via inhibitory or excitatory links. If cognition is indeed so implemented, then all mental transitions, including inferential ones, are associative in the implementation sense of the word. At same time, they might not be associative in the first sense. For more on different notions of association, see Mazzone 2021.

[46] Strictly speaking, the causal connection holds between token beliefs that encode the premises and the conclusions. The premises and the conclusions themselves are propositions, which are abstract objects and so can't be the relata of a causal relation. For brevity's sake, I will be using the terms "premise" and "conclusion" to talk about both propositions and mental states that encode them. It should be clear from the context which interpretation is intended.

taking is necessary for inferring, and instead look for different hallmarks. For example, Quilty-Dunn & Mandelbaum 2017 argue that a transition is inferential if and only if it is a result of mental processes that are sensitive to the logical form of the representations involved.

Each of these views has its own problems. If inference can be unconscious, accounts appealing to the subject's taking the premises to support the conclusion will not work. Some beliefs featuring in unconscious inferences (if such inferences exist) are, by definition, not accessible to the subject, so she can't be aware, or become aware, of why she has moved from the premises to the conclusion. Now, there are good reasons to think that inference can be unconscious, which puts pressure on the taking strategy (Bendana & Mandelbaum 2021).

On the other hand, the appeal to sensitivity to the logical form is problematic because not all the transitions that we would call "inferential" are transitions in virtue of logical form. Consider cases of semantic entailment, like the following: x is red, so x is colored. Quilty-Dunn and Mandelbaum have to say either that this transition is carried out in virtue of logical form (for example, by insisting that transitions like that always involve a tokening of the major premise "If x is red, then x is colored"), or that the transition is not inferential. Both options are problematic[47].

Siegel's RH faces none of these difficulties. In her view, inference is a distinctive kind of response to informational mental states. What makes it special is the fact that, unlike in other kinds of responding, the output of an inference epistemically depends on its input. As Siegel puts it, "… if you infer from poorly justified beliefs, or from experiences or intuitions that fail

[47] The claim that a transition from "x is red" to "x is colored" always involves tokening the major premise seems psychologically implausible: surely, sometimes one moves from the thought that apples are red to the thought that apples are colored without thinking that if apples are red, then apples are colored. Quilty-Dunn and Mandelbaum acknowledge the existence of such cases, but they deny that these are cases of inference. Rather, they take them to be instances of association: such transitions are carried out because concepts RED and COLORED become associated as a result of repeated inferences. Here intuitions diverge. One might accept their verdict. Or one might argue that the nature of mental processes is irrelevant for whether we should call such transitions "inferential". I discuss these competing approaches to defining inference in section 5.

to provide any justification… your conclusion will be poorly justified. *The hallmark* of inference is that the conclusions drawn by inferrers epistemically depend on the premises from which they are drawn." (Siegel 2019, p. 17, italics mine) This idea is attractive – so much so that the inferential belief is often defined as a belief whose justification depends on, or is derived from, other beliefs (Audi 2001, Lemos 2007). RH doesn't appeal to the subjects' taking the premise(s) to support the conclusion(s), so it can accommodate unconscious inference, unlike approaches invoking the taking state. Second, her account doesn't restrict inferences to transitions carried out by mental processes sensitive to the logical form of the representations involved in the transition, as Quilty-Dunn and Mandelbaum's account does[48].

Siegel uses different terminology to spell out the relation of epistemic dependence between the input and the output of an inference. Sometimes she talks about the output's responding to the reasons, or grounds, represented by the input (Siegel 2019, pp. 18-19). Sometimes, as in the quoted passage, she suggests that inference transfers justification, or warrant, from the input to the output. Finally, she says that an inference is rationally evaluable: a good inference responds to good reasons, which makes it rational, whereas a bad inference is irrational because it responds to bad reasons or responds to good reasons but do so in the wrong way. All these ways of describing epistemic dependence are, arguably, interconnected because a response that transfers justification is at the same time a response that is sensitive to reasons and is rational.

Siegel introduces her account of inference in what she calls "an exercise in illumination without analysis": instead of analyzing inference in terms of necessary or sufficient conditions, she

[48] One might argue that Siegel's RH and Quilty-Dunn and Mandelbaum's account don't directly compete with each other, since they adopt different approaches to defining "inference". Quilty-Dunn and Mandelbaum focus on psychological properties – namely, on the properties of mental processes underwriting inferential transitions, whereas Siegel is interested in the epistemic properties of such transitions. By doing so, they might end up talking about different phenomena. It doesn't mean, however, that the two accounts can't possibly be competitors. One can argue that the epistemic approach is preferrable because, say, it better captures our pre-theoretical intuitions about what inference is. If so, RH would have the edge over Quilty-Dunn and Mandelbaum's account. In section 5, I will say more about these two approaches to inference. For now, I will be relying on the epistemic approach, since it is the approach taken by Siegel.

contrasts it with other kinds of mental transitions to highlight the differences between them. So, what follows is a reconstruction of her account that is based on the examples and observations she makes.

The crucial role in her account is played by the notion of epistemic dependence. To repeat, the account boils down to the claim that a transition from state A to state B is inferential if and only if A causes B and B is epistemically dependent on A[49]. Here is an intuitive way of thinking about epistemic dependence: B's justificatory status depends on A's justificatory status if and only if A's justificatory status is relevant for B's justificatory status. I suggest that this relevance can be tested by the following counterfactuals:

(C) If A were not justified, B wouldn't be justified[50].

(C*) If A were justified, B would be justified[51].

---

[49] My reconstruction of RH emphasizes the role of epistemic dependence, which Siegel calls "*the hallmark* of inference" (Siegel 2019, p. 17, italics mine). Railton 2020 stresses another feature: the fact that "inference is a paradigm of person-level reasoning" (Siegel 2019, p. 15). On this interpretation, inferring is something that the agent does with the information she has. Appealing to an agentive aspect of inference restricts the scope of RH: for example, the account is not applicable to transitions between informational states in early visual cortex because one has no agency over such transitions. However, it is not clear whether the emphasis on the agentive aspect affects the dialectic of the paper. My counterexamples to RH feature associative transitions similar to a transition from a grandmother-related thought to a pie-related thought. Unlike states in early visual cortex, these thoughts are personal-level states. (There is another interpretation of "person-level reasoning". It holds that a transition from A to B counts as person-level reasoning only if the support relation between the input and the output is represented by a personal-level state. On this construal, associative transitions fail to be an instance of person-level reasoning. The problem with this reading is that many transitions that Siegel labels as "inferential" don't feature a representation of the support relation. If she took such a representation to be necessary for inference, her account would become an instance of the taking strategy.) But does it mean that associative transitions of this sort are something that the agent does? The agent's intentions don't seem to play any role in initiating them, and it might be even true that having an intention *not* to have a given associative transition has no power to prevent this transition from happening. However, Quilty-Dunn & Mandelbaum 2017 makes a strong case for thinking that, just like associative transitions, inferential transitions can be automatic. This suggests that the agentive aspect of inference should be compatible with automaticity. Similarly, the putative existence of unconscious inference speaks against too demanding an interpretation of the agentive aspect.

[50] The examples Siegel and I discuss involve transitions from one belief to another, so when I talk about a belief's being justified, I mean the justification one has for holding this belief, i.e., doxastic justification. Another clarification: what is at issue here is prima facie justification. This point will become important in section 5.

[51] One might worry that (C*) is too strong because it implies that the output of a bad inference doesn't epistemically depend on the input. Consider the error of affirming the consequent: *If A, then B; B; therefore, A*. Suppose that the two premises are justified. The conclusion, however, is not justified. Hence, this transition doesn't meet C* and thus doesn't count as inferential. This result seems problematic because affirming the consequent is still an inference, albeit a bad one. However, at this point we might not need to worry about cases of bad inference, since they might be parasitic on cases of good inference in the sense that bad inferences count as inferences only because they are sufficiently similar to good ones. Therefore, when constructing an account of inference, one should start with good cases, identify their commonalities and then extend the account to make it applicable to bad cases (see Grice 2001 and Wedgewood 2006 for this approach). For instance, one can argue that affirming the consequent counts as inference because it amounts to the application of an invalid logical schema that closely resembles the valid logical schema of *modus ponens*. And what makes *modus ponens* an inference is the epistemic dependence between

We have two kinds of cases to consider: cases where both the input and the output are justified, and cases where neither the input nor the output is justified. Let's start with the case where both the input and the output are justified. To determine whether the output epistemically depends on the input, one should apply (C): one should imagine a close possible world in which A is not justified, and consider what happens to the justificatory status of B. If B is not justified in this scenario, then B epistemically depends on A. In the case where neither the input nor the output is justified, one applies (C*), imagining a close possible world in which A is justified. If B becomes justified in this scenario, B epistemically depends on A.

In situations in which A and B are justified and A is the only state that causes B, (C) might seem unnecessary. After all, in such a scenario A is, as it were, the only place where B's justification might have come from[52]. However, (C) becomes useful in inferences whose input consists of several states, A_1, A_2, etc. How do we know whether, say, A_1 is relevant for B's justificatory status? We ask what would happen to B's status if A_1 were unjustified. If B's status changes from justified to unjustified[53], then B epistemically depends on A_1.

What about cases where A is justified and B is not, or *vice versa*? Here, the counterfactuals are unnecessary because it is immediately clear that A's justification is irrelevant for B's

---

the premises and the conclusion. Now, consider the following sequence of thoughts: *If A, then B; A; therefore, C*. Intuitively, this is not an inference, not even a bad one. Rather, it is a transition from some thoughts to another, completely unrelated thought. This intuition arises because the sequence doesn't resemble any valid inferential schema. To sum up, the proposed strategy holds that a bad inference counts as inference only because it is sufficiently similar to a good inference. Good inferences meet (C) and (C*) and thus exhibit epistemic dependence. More needs to be said about what counts as being sufficiently similar to a good inference, but this is the task for a different project.

[52] I assume that the belief B is not self-justifying, as, for example, is the belief that all triangles have three sides. Such self-justifying beliefs, arguably, don't epistemically depend on other beliefs. Accordingly, RH implies that they can't be inferred from other beliefs. This sounds correct: it seems infelicitous to say that one has inferred the belief that all triangles have three sides from another belief. In any event, inferential transitions discussed in the paper don't feature self-justifying beliefs, so we can put such beliefs aside.

[53] Justification comes in degrees. One can imagine a scenario where A_1 provides some degree of justification to B, A_2 adds to it, and so forth. In such cases, B epistemically depends on A_1 if the following condition holds: if A_1 were unjustified, B's degree of justification would decrease. Consequently, (C) and C*) should be slightly adjusted to account for such cases.

justification: one can obtain without the other. So, there is no epistemic dependence in such cases[54].

To illustrate: Consider the detective's reasoning in Whodunnit, which results in the belief that Smith is the murderer. In this case, the output epistemically depends on the input. If both the input and the output are justified, the relevant test for epistemic dependence is (C). Whodunnit passes this test, because, if the detective's premise beliefs were not justified, the conclusion would not be justified, either. Similarly, it is easy to see that Whodunnit passes the C* test when neither the input nor the output is not justified.

Or consider one of the examples of inference discussed in Siegel 2017 and Siegel 2019. In the Pepperoni case, a person eats a slice of pizza and suddenly realizes that the pizza is unappetizing. This belief arises in a response to different features of the pizza – it looks greasy, it is salty, it reminds the person of someone she dislikes[55], etc. Now, Siegel argues that the reasoning in Pepperoni qualifies as inference because the conclusion, namely the belief that the pizza is unappetizing, is epistemically dependent on the mental states (also beliefs, we can assume) that encode the features just mentioned. The tests for epistemic dependence I have proposed also applies to this example.

Now, consider a transition where the output doesn't depend on the input. A person entertains the belief that her grandmother is nearby. Suppose this belief is justified. Because the agent strongly associates her grandmother with apple pies, this thought leads her to believe that there

[54] I am grateful to the anonymous referee for pressing me to reconsider my characterization of epistemic dependence.

[55] The justificatory status of this belief (the belief that the pizza reminds the agent of someone she dislikes – let's call it "A") doesn't seem to affect the justificatory status of the conclusion that the pizza is unappetizing. This becomes clear when one applies (C) and (C*). One way to address this is point out that drawing the conclusion from A is a bad inference. As explained in footnote 9, such cases fall outside of the scope of (C) and (C*). Alternatively, one might take issue with Siegel's claim that the transition from A to the conclusion is an inference. It appears more akin to the transition from the thoughts if *A, then B,* and *A* to a completely unrelated thought *C* – something that shouldn't be labelled as "inference." Yet another way to account for this transition is to say that the example is underdescribed: the agent might also hold a belief that if a certain type of food reminds them of unpleasant people, the food is likely to be bad. If we incorporate this additional belief into the case, then this belief, belief A and the conclusion satisfy (C) and (C*).

is an apple pie nearby. The presence of her grandmother provides only weak justification for this second belief, rendering it unjustified. In this case, it is clear that the justificatory status of the output doesn't depend on the justificatory status of the input: if such dependence existed, the output would also be justified. This transition is associative and not inferential. The lack of epistemic dependence explains why.

So, from now on, I take RH to amount to the claim that a transition from A to B is inferential if and only if B is caused by A, and B epistemically depends on A, where epistemic dependence can be tested by C and C*.

Section 3: Associative inferential transitions?

Now that I have presented RH, I turn to one problem that it faces. I will argue that RH faces what I call "the demarcation problem": it doesn't adequately distinguish inferential from non-inferential transitions. The problem arises because the output states of some non-inferential transitions epistemically depend on their input states, which makes such transitions inferential in Siegel's view. More precisely, I will argue that some associative transitions exhibit epistemic dependency. Given that Siegel takes inference and association to be mutually exclusive, such cases pose a problem for her account.

Let's start by considering a hypothetical scenario that should pump the intuition that associative transitions exhibit epistemic dependence. Imagine organism O that lives in a simple environment. In this environment A-events are always followed by B-events. O has a capacity for associative learning and so it forms an association between A and B. O's mental states have propositional content of the sort A IS OVER THERE, and so on. One might therefore call A-states proto-beliefs. Suppose O's A-state is justified on a given occasion. Because A- and B-

states are linked, a tokening of A-state causes a tokening of B-state. Is the B-state justified in this situation?

It is tempting to say that it is justified because the transition is underwritten by a mechanism that (in these circumstances) reliably tracks the relationship between A-events and B-events that is relevant for B's justification. But this mechanism is associative, so prima facie we have an example of an associative transition exhibiting epistemic dependence. The transitions under discussion exhibit epistemic dependence because B-state's justification negatively depends on A-state's justification. We can stipulate that A-states are perceptual states, so normally they are justified.

Now, one might reply that RH is intended to cover transitions within the human mind, so it doesn't apply to O's transition. However, it is preferable to have an account that is equally applicable to non-human animals. Undoubtedly, non-human animals form associations; and we shouldn't rule out that they are capable of inferring (Kornblith 2012). Moreover, we could imagine similarly simple environments for human animals, i.e., environments in which A-events are always followed by B-events. Think of the child who is prima facie justified in thinking that the ice-cream van is at her front door as soon as she hears the ice-cream van's typical bell. In this scenario, we can stipulate that transitions in question are associative: a bell-sound belief is associatively followed by an ice-cream belief; the child doesn't reflect on whether the ice-cream belief is justified, and yet one still has a strong intuition that this belief is justified.

Let's now turn to an experiment suggesting that some associative transitions can exhibit epistemic dependence. Long-term memory is not a single faculty – rather, it is a host of systems that differ in their mechanisms and domains of operations (Squire 2004). One common

taxonomy of memory systems divides them into declarative and implicit memory. Declarative memory's representations can be consciously accessed, whereas implicit memory's representations can't. Implicit memory includes a system that underlies associative learning. Consider one task that arguably relies on this system – the weather prediction task (Knowlton et al., 1996) [56]. In this task, participants see various combinations of four different cards on each trial. Suppose these four cards are triangles, squares, circles and rectangles. Each card – regardless of which other cards are shown – is followed by the sun and the rain icons with a fixed frequency. For example, a card with triangles can be followed by the sun icon 75 per cent of the time. On every trial, participants first see a combination of these geometric shape cards and then are asked which weather icon will follow. At first, they have no information and so they guess; their performance is at the chance level. But with practice – after 40 trials or so – they give the correct response 70 per cent of the time.

Schematically, the cognitive processes that underwrite the performance on each trial can be represented as follows: The input of the process is a perceptual belief ("There is a card with triangles in front of me"), the output is a belief about the weather icon that will follow ("The sun icon will follow"). Now the question is, what processes underwrite this transition?

There are two competing descriptions of these processes: either the transition is underwritten by associations between geometric shape cards and weather icons, or it is mediated by beliefs about the stimuli contingencies ("triangle cards are often followed by the sun icon"). If the belief description is correct, then mental processes involved in the weather prediction task don't differ from ordinary instances of inference, in which one forms certain beliefs and uses them to arrive at other beliefs. So, if the weather prediction task is to constitute a counterexample to

[56] The first reason has to do with nature of the task: in this task, some stimuli are repeatedly followed by other stimuli, and it is exactly the circumstances in which one expects associative learning to take place. Second, people suffering from Parkinson's disease don't get better at this task: no matter how long they practice, their performance remains at the chance level (Knowlton et al., 1996). We know that Parkinson's disease damages neural structures called "basal ganglia" (Blandini et al. 2000), and basal ganglia are believed to be crucial for associative learning (Sheth et al. 2011).

RH – an example of *associative* transitions exhibiting epistemic dependence – the belief description must be ruled out.

The belief explanation is certainly an option that needs to be examined. One reason for this is the existence of propositional accounts of associative learning. According to them, associative learning depends on high-level cognitive processes that give rise to propositional knowledge, i.e., beliefs about co-occurring stimuli (Mitchell et al. 2009). Nonetheless, there are good reasons to reject this explanation.

To begin with, participants in this task have poor insight into the strategies they used. To give a simplified example: Suppose that triangle cards are followed by the sun icon on 75 per cent of the trials. Also suppose that the participant is very good at 'guessing' which icon follows triangle cards. The mediating-belief explanation holds that she forms the belief that the triangle shape is a good indicator of the sun icon, and uses it to predict the weather icon. However, if we ask her about the strategy she uses, she either won't be able to answer this question or will mention a strategy that she hasn't relied on: such a strategy can't account for her performance (Gluck et al. 2002). The belief explanation has troubles accounting for this poor insight because it is not clear why the participant can't report the relevant belief. By contrast, the associative explanation has an edge: according to it, the statistical relation between the triangle cards and the sun icon is not encoded by a representation with propositional content – instead, it is encoded by the strength of connections between the relevant concepts. The strength of connections is not accessible to the subject, which explains why she can't report it. Incidentally, the existence of such dissociation between performance on associative learning task and self-reports is one reason why some reject propositional accounts of associative learning (Baeyens et al. 2009).

The proponent of the belief explanation can reply that the beliefs in question are unconscious, but this move seems *ad hoc* unless we have reason to think that such beliefs are unconscious. Moreover, there are instances of associative learning taking place over stimuli that are not consciously represented (Pessiglione et al. 2008). Is it reasonable to think that one forms beliefs about unconscious stimuli?

If we rule out the belief explanation, we are left with the associative description of the task performance. The information about stimuli co-occurrences gets encoded in the strength of associative links between the relevant representations and is then used as follows: one sees a combination of geometric shapes; each shape is associated with some weather icon; and the stronger the associative links are (which is a function of a number of co-occurrences), the more likely it is that the representation of a given icon is activated. When it is activated, one forms the respective output belief ("The sun icon will appear").

Now, two final observations about the task. First, the output beliefs about weather icons are, arguably, justified after a certain number of trials. After all, participants get it right quite often, and it is no accident: the cognitive processes that underwrite the transition from the input to the output reliably track the information that is crucial for the successful performance, i.e., the information about the frequency with which a given shape is followed by a given icon.

This verdict about the justificatory status of output beliefs is intuitively plausible. It is not based on any particular account of justification, but an obvious candidate would be process reliabilism (Goldman 1979). Using the apparatus of this theory, one can argue that transitions from inputs to outputs in the weather prediction task constitute a reliable process type. This process type is belief-dependent because it takes as inputs perceptual beliefs about the cards with geometrical shapes. The process is conditionally reliable because it outputs a sufficiently high number of true beliefs only if its input beliefs are true. Indeed, if one were to hallucinate

a card with a geometric shape instead of having seen one, the output belief about which weather card would follow is likely to be false. Finally, one can add a requirement that the input beliefs in question must themselves be justified for the output to be justified. Overall, the reliabilist would agree that output beliefs in the task are justified, because the processes underlying task performance are conditionally reliable.

Second, output beliefs in the task are epistemically dependent on input beliefs. Recall that epistemic justification can be tested by (C) and (C*). These counterfactuals are applicable to input and output beliefs in the task. We can imagine that a participant has an unjustified quasi-perceptual belief about a card with a geometric shape (perhaps as a result of wishful thinking: the agent really wants the next card to be a triangle card and so she forms the respective belief without any good reason). This belief will cause an output belief (say, a belief that the sun card will appear). The intuitive verdict about the case is that the output belief is unjustified. After all, in this scenario one is out of touch with the information that is crucial for successfully performing the task, namely, the information about which geometric card is shown. By contrast, if we imagine that the input belief is justified, the output belief would be justified as well. If one wants to support the intuitive verdict by some account of justification, one can, again, think of reliabilism. It gives the same verdict about the task.

So, transitions between the input and the output in the weather prediction task exhibit epistemic dependence and therefore count as inferential on RH. At the same time, the transitions in question are associative. Given that Siegel takes inference and association to be mutually exclusive, here we have a counterexample to her view.

Putting cognitive science aside, the idea that associative transitions can sometimes output to justified beliefs can be supported by everyday examples. In fact, Siegel herself gives one such example: Imagine a child who stands next to a puddle and tries to figure out whether she can

jump over it (Siegel 2019). Siegel argues that the child's response to this perceptual situation is an inference because it has the hallmark of inferential transitions: epistemic dependency. Her belief that she can jump over the puddle is sensitive to reasons: it is justified, or rational, if it responds to good reasons, and is unjustified, or irrational, if it fails to respond to them. However, one can argue that in this case the decision-making is underwritten by reinforcement learning: in the past, some attempts to jump over a puddle of a similar size were successful, so they resulted in a positive experience and were reinforced, whereas some attempts led to the opposite result (Railton 2020). Suppose that the child succeeded in most of her attempts, so she believes that she can jump over this puddle. I think that many would be inclined to say that this belief is warranted even though it was arrived at in an associative fashion.

This is a good place to contrast my objection to RH with the one discussed in Railton 2020. As I said, Peter Railton points out that the decision-making in the puddle case might be underwritten by reinforcement learning. If that's the case, this scenario is *not* an instance of inference, argues Railton. The moral that he draws from this is that RH must say something about the underlying structure of given transitions because, as the puddle case shows, this structure is relevant for whether a transition is inferential.

My objection differs from Railton's in two respects. First, the central example I rely on is not the puddle case, but the weather prediction task. This is important because the puddle case is an imaginary scenario, so we don't know what's going on there internally. Specifically, we don't know whether the child's response is only mediated by reinforcement learning or whether, in addition, some of the child's beliefs play a role. It seems plausible that the reinforcement history might give rise to certain beliefs about puddles and these beliefs might in turn be used in decision making. So, one might argue that the puddle case doesn't constitute a counterexample to RH. By contrast, we have a better understanding of the processes underlying performance in the weather prediction task. As I argued before, there are good

reasons to reject the belief explanation of such performance. Therefore, one can't deal with this counterexample in the same way as with the puddle case.

Second, Railton and I draw different conclusions from counterexamples of this sort. He thinks that, to address them, RH must be furnished with some details about what kinds of mental processes constitute inference. I, by contrast, think that inference and association don't necessarily have to be mutually exclusive notions. I will defend this claim in Section 5. But before, let me address some objections to my counterexamples against RH.

Section 4: Objections and replies

Objection 1: Internalism and externalism

The argument from the previous section relied on an intuitive notion of justification. It is, however, clear that different accounts of justification would give different verdicts about the cases discussed above[57]. For example, I mentioned reliabilism and argued that its proponent should agree that associative transitions taking place in the weather prediction task exhibit epistemic dependence. So, to a first approximation, externalist accounts should be more friendly to the idea of epistemic dependence between the states of an associative transition, whereas most versions of epistemic internalism should be more skeptical about it. A standard formulation of internalism holds that an agent A's belief is justified only if the factors relevant to its justification are accessible to A, where "accessible" is often interpreted as introspectively accessible (Pryor 2001). Versions of internalism that subscribe to this claim are known under the label "accessibilism". Now, internalist accounts of this sort hold that one doesn't have

[57] RH as such is compatible with different accounts of justification. It cashes out inference in terms of partial epistemic dependence, which is, in turn, explicated by the following counterfactual: B epistemically depends on A just in case if A weren't justified, B wouldn't be justified. This counterfactual is silent on, for instance, whether such dependence requires awareness. If, for example, one is an internalist who thinks that epistemic dependence requires awareness, she can add the awareness condition to the account of epistemic dependence – no inconsistency would ensue.

justified beliefs in the discussed counterexamples to RH, since in these scenarios one has no introspective access to the factors that are relevant for the justification of these beliefs. For example, one doesn't have access to the links between the representations of the triangle cards and of the sun icons.

So, some internalists won't be persuaded by the argument of the previous section. They won't agree that the outputs of a mechanism whose workings are inaccessible to the subject are justified. Therefore, the argument is conditional: it will work only if one accepts certain views of justification. Defending a particular view of justification is beyond the scope of the paper, but I will mention just one consideration that favors externalism in the context of discussing Siegel's account.

Recall that RH can accommodate unconscious inference because epistemic dependence between the input and the output of a transition doesn't require awareness. In other words, one can inferentially respond to some informational state(s) without knowing what state(s) one is responding to. Siegel gives a few examples of such transitions (Siegel 2017, Siegel 2019). Moreover, a case for the existence of unconscious inference is supported by empirical data: the phenomenon of cognitive dissonance reduction arguably involves unconscious inferential transitions (Bendana & Mandelbaum 2021).

Now, a proponent of RH faces the following dilemma. On the one hand, if she accepts externalism, she can accommodate unconscious inference, but she is vulnerable to counterexamples like those discussed in the previous section. She would thus be forced to label certain associative transitions as "inferential". On the other hand, if she rejects externalism, she can handle these counterexamples, but she loses the ability to account for unconscious inference, since RH's explanation relies on the idea that unconscious mental states can exhibit

epistemic dependence. The upshot is that RH can address the counterexamples, but only at the cost of one of its key advantages.

But given the variety of internalist and externalist accounts, the dilemma might oversimplify matters. For example, mentalist internalism doesn't require that the subject have access to the mental states relevant to the justification (Conee & Feldman 2001). Its proponents could find a way out of the dilemma: they could hold both that some unconscious transitions are inferential and that associative transitions don't exhibit epistemic dependence. Let me unpack this claim. The examples of unconscious inference that are discussed in the relevant literature feature unconscious beliefs (Bendana & Mandelbaum 2021)[58]. Such beliefs are inaccessible to the subject *ex hypothesi*[59]; nonetheless, they are internal states of the subject, and the premise beliefs support the conclusion beliefs. So, mentalist internalism can account for epistemic dependence in unconscious inference: such inference features transitions between the agent's internal states, the input states epistemically support the output states, and so the output epistemically depends on the input. Associative transitions are different: even though associative links between representations are internal states of the subject, such links are not truth-evaluable (that is, they don't have propositional content), so they don't provide epistemic support for the output states of an associative transition. For example, the states that underwrite

[58] Mentalism defended in Conee & Feldman 2001 amounts to the claim that the only factors relevant for a belief's justification are some mental states of the agent holding this belief. Conee & Feldman 2001 doesn't mention unconscious inference, but their view leaves room for it. Consider the following example of unconscious inference: the agent has two unconscious beliefs: *A* and *if A then B*. These two states cause her to unconsciously believe that *B*. Suppose that the premise beliefs are justified. A mentalist can then argue that the output belief is justified as well because the contents *A* and *If A then B* epistemically support *B*, the agent believes that *A* and *If A then B*, and these beliefs are themselves justified. Mentalism can accommodate unconscious inference because the view doesn't require the factors relevant for justification to be consciously accessible to the agent – it only requires that such factors be her internal states. The existence of unconscious inference is a separate issue. For an argument in its favor, see Bendana and Mandelbaum 2021.

[59] Some might find the notion of unconscious belief to be incompatible with the standard usage of the word "belief". Even if it is true, this notion is compatible with the following influential picture of belief: believing that P consists in having a representation with the content P stored in the belief box. "The belief box" stands for whatever functional role that is played by beliefs. The exact functional role is a matter of controversy, but many agree that beliefs are sensitive to evidence, inferentially interact with other personal-level mental states and play action-guiding role (Schwitzgebel 2023). Arguably, a representation can play all these roles without its content being accessible to the agent. One can appeal to, say, self-deception to explain why the contents of certain beliefs are unconscious (Funkhouser & Barrett 2016).

a transition between the input and the output in the weather prediction task are associative links between concepts like TRIANGLE and SUN. The mentalist internalist would probably deny that such links provide sufficient justification for output beliefs[60].

To fully defend the idea that associative transitions exhibit epistemic dependence, one would need to make a case for an account of justification that leads to accepting the dilemma outlined above. However, the fact that some reasonable accounts of justification do face this dilemma strengthens my argument and highlights a significant challenge to Siegel's account. While it may not outright refute her position, it certainly raises a substantial and legitimate issue.

Objection 2: The description of the cognitive processes in the weather prediction task is a simplification

The weather prediction task is the central counterexample to RH, so a lot depends on whether my description of the cognitive processes involved in the task is correct. Recall that I argued that the transition from the input to the output belief is mediated by purely associative processes. Essentially, an input belief activates some associations, and one of those associations causes the output belief. One can argue that this description oversimplifies matters: more precisely, one can agree that associative processes do play a role, but then insist that those processes go in parallel with "normal" reasoning, which consists in evaluating the output of

[60] To better understand the difference between unconscious inference and associative transitions, let's contrast the transitions in the weather prediction task with the unconscious modus ponens mentioned earlier. The input states of the unconscious modus ponens have propositional contents (*A* and *If A, then B*) and these contents epistemically support the content of the output state, *B* (they logically entail it). By contrast, the input in the weather prediction task (say, the belief *There is a card with triangles in front of me*) doesn't support the output (say, the belief *The sun icon will appear*). What about the mental states mediating the transition from the input to the output – do they provide support for the output? The statistical relation between the triangle cards and the sun icon is not encoded by a representation with propositional content – instead, it is encoded by the strength of connections between the relevant concepts. In other words, the mediating states in question are associative links between the concepts TRIANGLE and SUN, and these links don't represent the proposition that cards with triangles are followed by sun icons. So, a mentalist would say that the output belief in the weather prediction task is not justified because the agent doesn't have mental states whose contents would support it. By contrast, the mental states possessed by the agent in the case of unconscious modus ponens support the output, which makes it justified. For more on the question of whether associations have propositional content, see Mazzone 2021.

the associative processes. If so, the output belief is a result of both associative and inferential processes. The task would thus become a less clear counterexample to Siegel's account.

I agree that my description is a simplification and that there are likely to be other processes going on in parallel. However, it is not clear that those processes are essential for epistemic dependence. We can imagine a participant who, while performing the task, has no thoughts apart from the input and the output beliefs. This person uncritically accepts a thought that pops into her head as a result of associative transitions. Compare her to somebody who critically examines their associative output before accepting it. Is such critical examination necessary for the resulting belief to be justified? It is not obvious that this is so. After all, we know that explicit reasoning plays no role in the task: participants' self-reports about the strategies they used are unreliable, declarative memory doesn't seem to be involved, and so on. So, explicit reasoning might not improve the performance. Why should we then think it is necessary for justification?

One might object to the claim that an uncritically accepted thought that just pops into one's head can be justified. Such thoughts, the objection goes, are very similar to deliverances of the reliable clairvoyance capacity from Bonjour's objection to reliabilism (Bonjour 1980). If such deliverances are not justified, the beliefs popping into one's head during the weather prediction task are not justified, either.

In response, let me discuss two cases. In both, the agent uncritically accepts a belief and yet, this belief seems justified. First, think of an average football player who has learned how to successfully dribble the ball past a defender, but only when the defender's legs are in a certain position. The player is not an expert, so she has just an intuitive understanding of the game. Now think of her thought that she can dribble the ball past the defender, when she notices that the defender's legs are in the mentioned position. Arguably, this belief is justified given what

she sees. Yet, she might not be able to provide any reason for this belief, since it is the result of pattern recognition learning and muscle memory. So, she might not know that the defender's position justifies the move that she is about to make.

Second, consider the experience of sitting in a crowded place like a bus and suddenly thinking that you are being watched. You quickly look up and lock eyes with a person nearby. One might explain this behavior in the following way: you unconsciously saw the person watching you, and this unconscious perception gave rise to the conscious belief that you are being watched (Berger 2014). Now, if one thinks that unconscious perception can justify beliefs, the belief under discussion is justified (Berger et al. 2018). Again, you are not in a position to provide reasons for this belief, but it doesn't prevent it from being justified.

These examples support the intuition that the agent might have a justified belief, even if she is not aware of the reasons on which the belief is based. Moreover, pattern recognition and muscle memory might heavily rely on associative transitions[61]. All this provides additional support to the claim that associative transitions might result in justified beliefs.

Section 5: Do inference and association have to be mutually exclusive kinds of mental processes?

Previously, I argued that RH faces what I called "the demarcation problem": it fails to distinguish inference from associative transitions. I discussed a few cases in which associative transitions yield justified beliefs and argued that RH classifies those transitions as inferential. This is a costly result for Siegel because she takes inference and association to be mutually

[61] I am grateful to the anonymous referee for drawing my attention to this point.

exclusive (Siegel 2019, p. 26). She is not alone in thinking that: Boghossian 2014 and Quilty-Dunn & Mandelbaum 2020 sharply distinguish between inferential and associative transitions.

But why should we accept this sharp division? In this section, I offer a way out for RH that challenges the mutually exclusive view of inference and association, and discuss the motivation behind this move. I will argue that one should simply accept the existence of associative inferential transitions. This move may strike some as rather bold, since many discussions of inference in philosophy and cognitive science tacitly presuppose the sharp division between inference and association. Here, I provide some reasons that will make the move not only palatable but also interesting and fruitful.

As I have previously discussed, there are at least two types of approaches to spelling out the notion of inference. Either one defines inference as a kind of transition brought about by certain mental processes (e.g. processes that are sensitive to the logical structure of representations, see Quilty-Dunn & Mandelbaum 2017), or one spells out this notion in epistemic terms (this is the route taken by Siegel, see also Lemos 2007, Audi 2001)[62]. Both approaches face difficulties. This paper focuses on the difficulties faced by the epistemic approach: I argued that it leads to the demarcation problem. If the mark of inference is epistemic dependence, then some associative transitions are inferential.

[62] Epistemic and psychological properties are sometimes interconnected. To illustrate: some internalists argue that state A can justify state B only if the agent is aware of A. Combining this account of justification with RH yields an account of inference that requires the input state of a transition's to be conscious. However, the connection between the cognitive and the epistemic can be looser. Consider process reliabilism, which holds that a mental state is justified if and only if it is an output of a reliable process, where "reliable" roughly means producing a high ratio of true beliefs. Many processes fit with this description: perception, memory, deductive reasoning and so on. Cognitively speaking, they are very different from each other. But from a reliabilist's point of view, the justificatory status of their outputs might be equivalent. To sum up, different accounts of justification have different consequences for what counts as epistemic dependence, and this in turn has consequences for what kind of mental processes can implement epistemic dependence. As a result, the relationship between the cognitive and the epistemic might be more or less tight. One can also develop an account distinguishing different kinds of inference, ranging from more automatic and less sophisticated transitions to reflective reasoning, and argue that they differ with respect to the justificatory status of their outputs. (Mazzone 2021). Given this interplay between the cognitive and the epistemic, some accounts of inference might be couched in both psychological and epistemic vocabulary. I thank the anonymous referee for pressing me to elaborate on the relationship between epistemic and psychological properties.

But what generates this intuition about epistemic dependence exhibited by some associative transitions? Several factors might be at play. To begin with, the notion of association is spelled out in psychological terms and is thus silent on whether such transitions exhibit epistemic dependence. Recall that a transition from representation A to representation B counts as associative just in case the stimuli represented by A and B co-occurred in one's experience and, as a result of such co-occurrence, a tokening of one representation causes a tokening of the other. This characterization is compatible with B's justificatory status being dependent of A's justificatory status in some circumstances. For example, associative transitions might result in justified beliefs in situations similar to the weather prediction task[63].

Second, our intuitions about epistemic dependence are sensitive to, among other things, our intuitions about the justificatory status of a mental state. The latter is in turn influenced by different factors. Clearly, one factor is the nature of mental processes underwriting a given transition. But no less important is, to put it somewhat vaguely, the situation in which mental processes are embedded. For example, the justificatory status of perceptual beliefs is, on some views on higher-order evidence, affected by beliefs about the reliability of this faculty. The latter are not causally implicated in perception, but, arguably, are relevant for the justificatory status of its output (Bradley 2019). That is, perceptual beliefs are not *prima facie* justified if

[63] One might argue that any associative transition provides some degree of justification for the output belief. After all, one comes to associate stimuli A and B if A and B co-occurred in one's experience. Given this statistical relationship between A and B, the presence of A might provide a certain degree of justification to the belief that B is present. Going back to the grandmother example: the fact that one's grandmother is nearby makes one think that an apple pie is nearby, and this latter belief might not be completely unwarranted. While this reasoning has some plausibility when one thinks about associations tracking co-occurring stimuli, it becomes much less plausible when applied to associations formed solely as a result of co-activation of concepts. One can form an association between concept A and concept B even if the things represented by A and B never co-occurred in one's experience. Say, a person has been repeatedly told that people of race R are dangerous. As a result, she has an association between the concepts R and DANGEROUS. We can stipulate that the claim about R is false and that the source of this claim is not trustworthy. Imagine that the agent encounters a member of R and, because of the association, forms the belief that this person is dangerous. In this scenario, it seems problematic to say that the association in question provides justification to the resulting belief.

one has reasons to think that one's perception is unreliable. So, the same perceptual processes might result in justified beliefs in one situation and fail to do so in another, depending on the background beliefs that one has.

The same conclusion follows from the views of perceptual warrant defended by Wright and Coliva (Wright 2004, Coliva 2015). Wright defends a view according to which perceptual experiences justify perceptual beliefs only if the agent is rationally entitled to accept some background presuppositions. Suppose that the proposition that one's perceptual mechanisms function properly (P) is among them. On Wright's account, being entitled to accept P requires the absence of the belief that not-P. Accordingly, if one has evidence against P and so comes to disbelieve P, one's perceptual beliefs are not even prima facie justified. Coliva defends a similar view of perceptual warrant: she agrees with Wright that perceptual experiences provide justification only if one accepts some presuppositions (she calls them "assumptions"), but, unlike Wright, she argues that such presuppositions don't have to be warranted themselves.

Furthermore, our intuitions about the justificatory status of a mental state might be sensitive to factors that are not even part of one's cognitive system. Suppose that the agent doesn't make use of readily available higher-order evidence about the unreliability of perceptual processes. Arguably, such external evidence makes a difference for the justificatory status of the output of a given process. Consider an example along these lines (Gibbons 2006): One is cooking a cream cheese omelette for breakfast. The only ingredient that is left to add is cheese. The agent thinks that the cheese is in the fridge, and her thinking is based on good reasons: she remembers that it was there yesterday in the evening, she has a justified belief that her flatmate doesn't usually eat breakfast, and so on. As a result, she believes that the cheese omelette will soon be ready. Is her belief justified? It seems so, but here is the twist: there is a note on the fridge door reading, 'We are out of cream cheese.' In this household, it is customary to leave notes of this sort on the fridge. Gibbons argues that the belief in question is not justified because the agent

failed to take into account the evidence that was readily available to her. By contrast, in a nearby possible world in which the subject has exactly the same mental states but in which there is no note on the fridge, her beliefs are justified. In other words, we have two cases where the very same reasoning process outputs beliefs with different justificatory statuses. What accounts for this variance is not an intrinsic difference between the subjects (they are stipulated to be intrinsically indistinguishable) but something extrinsic to them: the note on the fridge that the subject missed in one case and that was absent in the other.

These examples show that factors extrinsic to a mental transition are relevant for our judgments about whether its output is justified. By the same token, the presence of certain extrinsic factors can influence what we think about the justificatory status of the output of an associative transition. The weather prediction task is one example in which one might have

the intuition that the output of an associative transition is justified, presumably because this task provides the perfect environment for the employment of associative learning mechanisms. And the same, mutatis mutandis, applies to the other counterexamples from section 3. But if one accepts that associative transitions sometimes result in justified beliefs, one can then argue that these transitions exhibit epistemic dependence.

Where does this leave us? If one is attracted to the claim that the mark of inferential transitions is epistemic dependence, one should be ready to reject the assumption that inference and association are mutually exclusive. In other words, if, in spelling out the notion of inferential transitions, one cares not about the nature of the mental processes that underwrite such transitions, but about their epistemic properties, then one should be ready to accept the existence of inferential associative transitions. Now, Siegel's account aims to do exactly that, i.e., to spell out the notion of inferential transitions in epistemic terms while abstracting away from the structure of mental processes underwriting them. Therefore, she shouldn't be

committed to the claim that inference and association are mutually exclusive. If this claim is rejected, we can restore the value of RH.

Arguably, the rejection of the mutually exclusive claim is warranted only if we opt for the epistemic approach to the notion of inference, and I have provided no argument to show that such an epistemic view is better than the structure-of-mental-processes approach. So, a more conservative conclusion should be formulated along the following lines. If the explanatory project is to spell out, say, the difference between System 1 and System 2 processes, then the mental-processes approach is preferable: System 1 processes are associative and don't overlap with System 2 processes, so if System 2 processes are inferential, inference should be sharply distinguished from association. If, on the other hand, the explanatory project has a more epistemological flavor, a notion of inference that captures the idea of epistemic dependence appears to be much more appropriate. Such an epistemic approach would inevitably lump together different kinds of mental processes, including associative processes, i.e., this approach wouldn't be able to solve the demarcation problem. However, the failure to solve this problem is not a reason for concern because the aim of the epistemic approach is not to distinguish between inferential and associative transitions; rather, the aim is to pick out a class of transitions that have some interesting epistemic properties in common. If some associative transitions happen to have such epistemic properties, so be it.

Conclusion

I have argued that RH, the account of inference recently proposed by Susanna Siegel, faces what I have called "the demarcation problem": it can't distinguish between inferential and non-inferential transitions. RH is based on the plausible idea that the hallmark of an inferential transition is epistemic dependence of its output on its input. I have argued that some associative

transitions exhibit epistemic dependence characteristic of inference. For this reason, RH would classify such transitions as inferential, but this is a problematic verdict because Siegel and many others take inference and association to be mutually exclusive. I have suggested one way in which RH can accommodate these cases. The solution is to reject the assumption that inference and association are mutually exclusive kinds of mental transitions. This move is motivated by the observation that there are at least two approaches to explicating inference: one approach spells it out in terms of mental processes underwriting a given transition, whereas the other does so in epistemic terms. If one favors the second approach, one should be ready to accept that the demarcation problem can't be solved, since there might be no sharp boundary between inference defined in epistemic terms and, say, associative transitions. However, this shouldn't worry the proponent of the epistemic approach, since the aim of this approach is not to distinguish inferential transitions from associative transitions, but rather to pick out a class of transitions that have some interesting epistemic properties in common. So, rejecting the assumption that inference and association are mutually exclusive becomes a warranted move.

Declarations

Ethical Approval

Not applicable.

Funding

Research for this article was supported by MICIU/AEI/ 10.13039/501100011033 and by ERDF A way of making Europe, under grant agreement PID2021-124100NB-I00, by AGAUR, under grant agreement 2021-SGR-00276, and by Grant CEX2021-001169-M funded by MICIU/AEI/10.13039/501100011033. The author reports that there are no competing interests to declare.

Authors' contributions

Ilia Patronnikov

Acknowledgments

I thank Sven Rosenkranz, Michele Palmira and Daniel Gregory for their helpful comments on some of the drafts of this paper. I am also grateful to the audiences at the conferences where this work was presented: 29th SIUCC conference (University of Barcelona), 3rd joint conference of the ESPP and SPP (University of Milan), 2022's SOPhiA conference (University of Salzburg), 6th Philosophy of Language and Mind conference (University of Warsaw), 6th congress of the Spanish Society for Analytic Philosophy (University of Santiago de Compostela), the Epistemology Reading Group (LOGOS) and the Graduate Reading Group (LOGOS). Finally, I owe a lot to Pepa Toribio for her sharp comments and invaluable help.

References

Audi, R. (2001). The architecture of reason: The structure and substance of rationality. Oxford University Press.

Baeyens, F., Vansteenwegen, D., & Hermans, D. (2009). Associative learning requires associations, not propositions. *Behavioral and Brain Sciences*, *32*(2), 198–199. https://doi.org/10.1017/s0140525x09000867

Bendana, J., & Mandelbaum, E. (2021). The fragmentation of belief. In Borgoni, C., Kindermann, D., & Onofri, A (Eds.). The fragmented mind. (Oxford University Press), 78-108.

Berger, J. (2014). Mental states, conscious and nonconscious. *Philosophy Compass*, *9*(6), 392–401. https://doi.org/10.1111/phc3.12140

Berger, J., Nanay, B., & Quilty-Dunn, J. (2018). Unconscious perceptual justification. *Inquiry*, *61*(5–6), 569–589. https://doi.org/10.1080/0020174x.2018.1432413

Blandini, F., Nappi, G., Tassorelli, C., & Martignoni, E. (2000). Functional changes of the basal ganglia circuitry in parkinson's disease. Progress in Neurobiology, 62(1), 63–88. https://doi.org/10.1016/s0301-0082(99)00067-2

Bradley, D. (2019). Are ther indefeasible epistemic rules? Philosophers' Imprint, 19(3), 1-19

Boghossian, P. (2014). What is inference? Philosophical Studies, 169(1), 1–18. https://doi.org/10.1007/s11098-012-9903-x

Boghossian, P. (2018). Delimiting the boundaries of inference. Philosophical Issues, 28(1), 55–69. https://doi.org/10.1111/phis.12115

Bonjour, L. (1980). Externalist Theories of Empirical Knowledge. *Midwest Studies in Philosophy*, 5: 53–73. doi:10.1111/j.1475-4975.1980.tb00396.x

Bouton, M. (2016). Learning and behavior: A contemporary synthesis. Sinauer Associates.

Coliva, A. (2015). Extended rationality: A hinge epistemology. Palgrave Macmillan.

Conee, E., & Feldman, R. (1998). The generality problem for reliablilism . Philosophical Studies, 89(1), 1–29. https://doi.org/10.1023/a:1004243308503

Conee, E. & Feldman, R. (2001). "Internalism Defended." *American Philosophical Quarterly*, 38(1): 1–18

Frankish, K., & Evans, J. (2009). In Two Minds: Dual processes and beyong. Oxford University Press.

Funkhouser, E., & Barrett, D. (2016). Robust, unconscious self-deception: Strategic and flexible. *Philosophical Psychology*, *29*(5), 682–696. https://doi.org/10.1080/09515089.2015.1134769

Gibbons, J. (2006). Access externalism. Mind, 115(457), 19–39. https://doi.org/10.1093/mind/fzl019

Goldman, A. I. (1979). What is justified belief? Justification and Knowledge, 1–23. https://doi.org/10.1007/978-94-009-9493-5_1

Gluck, M. A., Shohamy, D., & Myers, C. (2002). How do people solve the "Weather prediction" task?: Individual variability in strategies for probabilistic category learning. Learning & Memory, 9(6), 408–418. https://doi.org/10.1101/lm.45202

Grice, P. (2001). Aspects of reason. Oxford University Press.

Kahneman, D. (2011). Thinking, fast and slow. Farrar, Straus and Giroux.

Kamin, L. (1969). Predictability, surprise, attention, and conditioning. in B. Campbell and R. Church (eds.), Punishment and aversive behavior (New York: Appleton-Century-Crofts), 279–296.

Knowlton, B. J., Mangels, J. A., & Squire, L. R. (1996). A neostriatal habit learning system in humans. Science, 273(5280), 1399–1402. https://doi.org/10.1126/science.273.5280.1399

Kornblith, H. (2012). On reflection. Oxford University Press.

Lemos, N. (2007). An introduction to the theory of knowledge. Cambridge University Press.

Mazzone, M. (2021). An associative account of inferences. Rivista internazionale di filosofia e psicologia

Mitchell, C. J., De Houwer, J., & Lovibond, P. F. (2009). The propositional nature of human associative learning. *Behavioral and Brain Sciences*, *32*(2), 183–198. https://doi.org/10.1017/s0140525x09000855

Pessiglione, M., Schmidt, L., Palminteri, S. et al. (2008). Reward processing and conscious awareness. In: Delgado MR, Phelps EA, Robbins TW (eds), Decision Making, Affect, and Learning: Attention and Performance XXIII Oxford/New York: OUP, 2011, 329–48.

Pryor, J. (2001). Highlights of recent epistemology. *The British Journal for the Philosophy of Science*, *52*(1), 95–124. https://doi.org/10.1093/bjps/52.1.95

Quilty-Dunn, J., & Mandelbaum, E. (2017). Inferential transitions. Australasian Journal of Philosophy, 96(3), 532–547. https://doi.org/10.1080/00048402.2017.1358754

Railton, P. (2020). Comment on Susanna Siegel, the rationality of perception. Philosophy and Phenomenological Research, 101(3), 735–754. https://doi.org/10.1111/phpr.12735

Rescorla, R. A. (1988). Pavlovian conditioning: It's not what you think it is. American Psychologist, 43(3), 151–160. https://doi.org/10.1037/0003-066x.43.3.151

Schwitzgebel, E. (2023) *Belief*, [Online], Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/entries/belief/ [10 January 2025]

Sheth, S. A., Abuelem, T., Gale, J. T., & Eskandar, E. N. (2011). Basal ganglia neurons dynamically facilitate exploration during associative learning. The Journal of Neuroscience, 31(13), 4878–4885. https://doi.org/10.1523/jneurosci.3658-10.2011

Siegel, S. (2017). The rationality of perception. Oxford University Press.

Siegel, S. (2019). Inference without Reckoning. in Reasoning: Essays on Theoretical and Practical Thinking, ed. M. Balcerak-Jackson and B. Balcerak-Jackson, Oxford: Oxford University Press.

Squire, L. R. (2004). Memory systems of the brain: A brief history and current perspective. Neurobiology of Learning and Memory, 82(3), 171–177. https://doi.org/10.1016/j.nlm.2004.06.005

Wedgwood, R. (2006). The normative force of reasoning. *Noûs*, *40*(4), 660–686. https://doi.org/10.1111/j.1468-0068.2006.00628.x

Wright, C. (2004). Warrant for nothing (and foundations for free)? Aristotelian Society Supplementary Volume, 78(1), 167–212. https://doi.org/10.1111/j.0309-7013.2004.00121.x

Chapter 4: Conclusion

Section 1: Results

In this thesis, I have examined the nature of implicit attitudes (IAs), focusing on the question of what kind of mental states they are. Among the many competing accounts, I have defended the belief view: the claim that IAs are beliefs. I have identified two types of arguments supporting this view. The first type appeals to empirical evidence suggesting that IAs exhibit characteristic features of beliefs, concluding that the belief view offers the best explanation of this evidence. The second type argues that disanalogies between IAs and paradigmatic beliefs are not sufficient to disqualify IAs from being beliefs.

My main contribution to the debate belongs to the second type of arguments. It focuses on the asymmetry between IAs and ordinary beliefs with respect to self-knowledge. Arguably, people are typically in a good epistemic position with respect to their beliefs: if one believes that P, one is usually in a position to know (or at least to have the justified belief) that one believes that P. By contrast, IAs are opaque: one's epistemic position vis-à-vis such attitudes is poor. But if IAs are beliefs, how could this be the case? I have called this objection "the self-knowledge objection" to the belief view. I have argued that IAs are not typically accompanied by the relevant judgments and, since judgment is a primary route to self-knowledge of belief, its absence explains the epistemic opacity of IAs.

I have also clarified a notion that plays a key role in arguments for the belief view – the notion of inference – thereby strengthening the conceptual foundations of the debate over IAs. I have identified two approaches to characterizing inference – the mental-processes approach and epistemic approach – and argued that only the mental-processes approach can support the arguments in question.

Section 2: Discussion

The findings presented raise several important theoretical issues and point toward promising avenues for further research.

First, the self-knowledge objection to the belief view brings into focus a distinctive feature of IAs: if they are beliefs, they are beliefs that are not accompanied by the relevant judgments. This invites further reflection on the relationship between believing and judging. Some accounts of belief hold that it is a conceptual truth that believing entails judging. If this conceptual analysis is correct, it prompts a broader methodological question: To what extent should our theorizing about belief be constrained by the folk concept?

This question is at the heart of a broader debate between common-sense functionalism and psychofunctionalism. According to the former, the functional role of mental states must be derived from the folk psychological concept. According to the latter, folk concepts are the starting point for theorizing but may be revised in light of empirical discoveries. Given the distinctive functional profile of IAs – partly belief-like, partly divergent – common-sense functionalists are likely to reject the belief view, whereas psychofunctionalists may find it plausible. Thus, the viability of the belief view is bound up with this deeper debate about the methodology of theorizing about mental kinds.

Another interesting feature of IAs is their epistemic deficiency – the fact that such attitudes are often evidentially problematic by one's own lights. This property may be relevant to the fragmentation approach to belief storage. If the belief view is correct, a person might have plainly inconsistent explicit and implicit beliefs about a certain social group. For example, she might believe that members of a certain race are less intelligent than others, and that they are equally intelligent. This raises the question of how such inconsistent beliefs are maintained.

Proponents of fragmentation speculate that they are stored in different fragments. One might wonder why they are so stored.

IAs' epistemic deficiency may provide an explanation. Here is a sketch of how it might go. Suppose one has an implicit belief P and an explicit belief Q, such that content P is evidentially problematic by one's own light and content Q aligns with one's evidence. Implicit content P, being evidentially unsupported, tends not to be retrieved in deliberative contexts, where its inferential defeat is apparent. By contrast, content Q keeps being brought up. Repetition of this suppression may lead to content P and content Q being stored in different cognitive fragments.

The opacity of IAs understood as beliefs also invites further exploration of the conditions under which self-knowledge of belief is achievable. I have argued that one important route to self-knowledge of belief is blocked for IAs. This argument, however, doesn't rule out that one can access one's IAs in some other way. For instance, one can argue that they can be known via inference from behavior and other mental states. If this method of self-knowledge is available for implicit beliefs, one might wonder why they are opaque.

I have speculated that this worry can be addressed by pointing to several factors. For one thing, implicit beliefs may be subject to motivated reasoning. If people want to preserve their positive self-image, they would be motivated *not* to discover certain facts about themselves. For example, they would be motivated to ignore manifestations of their racist beliefs. This explains how one's implicit beliefs may go undetected despite the abundance of evidence.

For another thing, implicit beliefs are an obvious target of a psychological immune system. A psychological immune system is hypothesized to be a set of mechanisms whose function consists in ameliorating the experience of negative affect. The realization that one harbors implicit beliefs are likely to cause such affect: if one is a committed egalitarian, knowing that one harbors racists beliefs may hurt. The mechanisms underwriting a psychological immune

system may work towards protecting one from the exposure to information leading to such hurtful realization. Accordingly, these mechanisms can explain poor inferential access to implicit beliefs.

Overall, the connection between IAs and the mentioned approaches and theories deserves further investigation.

My discussion of putative associative inferential transitions has resulted in the counter-intuitive conclusion that some associative transitions may result in justified beliefs. One may wonder about the scope of this claim. Clearly, situations in which associative transitions exhibit epistemic dependence are limited, and if one adopts the strategy of believing any thought that comes to their mind in an associative fashion, they will end up having a lot of false and unjustified beliefs. So, my claim about the inferential nature of some associative transitions should be strengthened by developing an account that allows to distinguish between circumstances in which associative process confer justification from circumstances in which they don't.

As the relevant discussion makes clear, the claim about associative inferential transitions is tied to the notion of epistemic justification. Different accounts of justification have different implications for plausibility of this claim. I have pointed out that externalist accounts are more sympathetic to this idea than internalist accounts. Cashing out and defending an account of justification compatible with the existence of associative inferential transition is a worthwhile project.

Finally, questions arise about the epistemic credentials of beliefs grounded in unconscious mental states. In cases such as the weather prediction task, output beliefs appear justified even though the agent lacks access to the mental states conferring this epistemic status. This raises several puzzling philosophical questions: How can unconscious mental states confer

justification? Does such justification differ from justification conferred by conscious mental states? Again, these issues warrant further investigation.

Section 3: Summary

This thesis tackles the question of what kind of mental states implicit attitudes (IA) are. In Chapter 1, I have outlined the phenomenon of IAs and the theoretical challenges they raise. After introducing the empirical methods used to study IAs, I have sketched the landscape of the views about the nature of IAs and examined evidence supporting one account – the account holding that IAs are beliefs. I have also explained why competing accounts have troubles accommodating this evidence and discussed some other advantages of the belief view. The remainder of Chapter 1 introduces the issues to be discussed in Chapter 2 and Chapter 3 and outlines the argumentative strategy of these chapters.

In Chapter 2, I have addressed what I call "the self-knowledge objection" to the belief view. It amounts to the following: People are typically in a good epistemic position with respect to their beliefs, they are often unaware of their IAs. But if IAs are beliefs, how can this asymmetry be explained? I argue that IAs constitute a special kind of belief – beliefs not accompanied by the relevant judgments and that judging is an important route to knowing one's beliefs. This route is blocked for IAs, which explains their opaqueness – one's bad epistemic position with respect to such attitudes. I have argued that the lack of alignment between believing and judging in the case of IAs is explained by epistemic deficiency – the fact that such attitudes are often evidentially problematic by one's own lights. I have outlined an account of belief that can accommodate these features – epistemic deficiency and the lack of alignment with the relevant judgments – of IAs.

In Chapter 3, I have examined the notion of inference. It is relevant to the debate on the nature of IAs because an important argument for the belief view holds that IAs feature in inferential transitions, thus indicating that they are beliefs. To assess this argument, we must clarify what inference is. Susanna Siegel's Response Hypothesis provides a promising answer to this question. Her account takes inferring to be a special kind responding to mental states in which the output mental state epistemically depends on the input. I have argued that her account is compatible with some associative transitions being inferential and offered several scenarios supporting this claim. This outcome is problematic for Siegel, since she takes inference and association to be mutually exclusive. I have suggested that she should abandon this assumption and provide a rationale for this move. I have argued that there are two approaches to understanding inference. On one view, which I will call *"*the mental-processes approach*"*, inference is defined in terms of the psychological structure underwriting transitions between mental states. By contrast, the other approach – let's call it "epistemic" – spells out the notion of inferential transition in epistemic terms while abstracting away from the structure of the mental processes underwriting them. Siegel's account exemplifies the epistemic approach, and since this approach is silent on the psychological structure of inferential transitions, her account should allow for the existence of associative inferential transitions, provided that such transitions exhibit epistemic dependence. This shows that epistemic approach allows for the possibility that states with associative structure feature in inferential transitions. Accordingly, the fact that IAs participate in inference-like patterns does not entail that they are beliefs. The upshot is that the force of the inference-based argument for the belief view depends crucially on which of the two approaches one adopts.