

Intuitive and Deliberative Behavior

The Role of Cognition in Sociological Models of Action

Sascha Grehl

Thesis for the degree of Doctor Philosophiae (dr. philos.)
University of Bergen, Norway
2024

UNIVERSITY OF BERGEN



Intuitive and Deliberative Behavior

The Role of Cognition in Sociological Models of Action

Sascha Grehl



Thesis for the degree of Doctor Philosophiae (dr. philos.)
at the University of Bergen

Date of defense: 22.03.2024

© Copyright Sascha Grehl

The material in this publication is covered by the provisions of the Copyright Act.

Year: 2024

Title: Intuitive and Deliberative Behavior

Name: Sascha Grehl

Print: Skipnes Kommunikasjon / University of Bergen

Scientific Environment

This doctoral project was developed within the research project “Foundations of game-theoretic solution concepts” (“Grundlagen spieltheoretischer Lösungskonzepte”), founded by the German Research Foundation from 2013 until 2019 (Reference number: TU 409/1-1 and TU 409/1-2) at the Leipzig University, Germany. During the doctoral project I have been formally employed at the Institute of Sociology in Leipzig. My supervisor during the project has been Dr. Andreas Tutić at the Department of Sociology. When Dr. Andreas Tutić received an invitation to hold the position of Associate Professor at the University of Bergen, our collaborative efforts continued without interruption. We maintained regular communication to discuss research progress and plan future collaborative work.

Acknowledgments

I would like to express my sincere gratitude to my supervisor, Andreas Tutić, for his invaluable guidance, support, and mentorship throughout the entire research process. His extensive knowledge, critical insights, and unwavering commitment to my academic and personal growth have been instrumental in shaping this thesis.

I would also like to thank the Department of Sociology in Bergen for their warm hospitality and support, which has made this dissertation possible.

Furthermore, I am indebted to the "Glas und Bohne" café for providing a conducive and inspiring environment where I could work on this thesis. The warm and inviting atmosphere, along with the delicious coffee, provided the perfect setting for me to focus and engage in productive work.

Lastly, I would like to express my deepest appreciation to my spouse and my two children, whose unending love, encouragement, and understanding have sustained me throughout this challenging academic journey. Their constant support and belief in my abilities have been a constant source of inspiration and motivation, and I am truly blessed to have them in my life. Without their love and support, this achievement would not have been possible.

Leipzig, July, 2023.

Sascha Grehl

Abstract

This thesis explores the role of cognition in shaping human behavior. Based on a dissatisfaction with the overly simplistic assumptions about human cognition prevalent in many contemporary sociological theories, this research seeks to contribute to the development and assessment of more sophisticated theories of action that more accurately capture the realities of the human mind. The argument presented in this thesis is that a more nuanced understanding of human cognition is not merely an intellectual exercise, but an indispensable prerequisite for achieving more precise and profound sociological explanations, while at the same time providing new and socially relevant predictions. For practical sociological research, however, it is crucial to ensure that such enriched theories remain applicable without becoming mired in unnecessary complexity. To address these objectives, this thesis presents six peer-reviewed contributions, linked by a framing introduction, that cover the following three distinct areas of inquiry:

First, the thesis commences with a critique of Rational Choice Theory (RCT), which is chosen as the starting point for this investigation because it serves as a typical example of action theories that rely on simplifying assumptions about the human mind. It then presents two laboratory experiments designed to gather more data about human cognition that may help refine existing or develop new theories of action. These experiments focus on the cognitive ability to engage in iterated reasoning, which is a critical prerequisite for many solution concepts used by RCT to predict behavior. The studies show that the human capacity for iterated reasoning is very limited and that there is little variation among individuals, providing a sharp contrast to the assumptions of RCT, which typically posit un-

limited capacities in this regard. As a consequence, the observed behavior does not match the predictions of RCT, underscoring the importance of accounting for cognitive limitations. Furthermore, the data suggest that in addition to cognitive ability, the individual reasoning style, i.e., how one utilizes one's cognitive abilities, has a significant impact on decision making that appears to be even stronger than that of cognitive ability.

Second, expanding upon the preceding insights, this thesis then places more emphasis on the study of reasoning styles. In this context, two existing theoretical approaches that address the distinction between cognitive reasoning styles are examined: The Dual Process Perspective (DPP) and the Status Characteristics Theory (SCT). The DPP is a general theoretical framework that distinguishes between intuitive and deliberative reasoning, outlining their combined effects on behavior. In contrast, SCT can be viewed as a specialized application of the DPP, focusing on how status cues intuitively influence behavior in the absence of deliberate intervention. Following the theoretical explication of both approaches, specific hypotheses are derived. For the DPP, these hypotheses focus on the conditions under which either intuitive or deliberative reasoning might exert greater influence and the possible consequences of such influence. SCT is used to predict which status cues are likely to influence behavior and which cues are not. To validate these hypotheses, two additional laboratory experiments are conducted. The results generally corroborate the main tenets of both DPP and SCT, underscoring the value of incorporating cognitive facets into sociological models of action.

Third, the thesis examines whether the DPP can be applied to sociological research outside the laboratory. To this end, the thesis presents empirical evidence from a non-reactive field experiment and an online survey. The field experiment serves as a litmus test for the external validity of the DPP by applying its principles in a real-world context without participants being aware of their involvement in a study. The online survey evaluates the applicability of the core concepts of the DPP within the framework of standard sociological survey techniques. Both studies confirm the usefulness of the DPP in real-world scenarios and its compatibility with common online survey formats.

In conclusion, this thesis provides clear evidence challenging the notion that humans possess unlimited cognitive capacities. More importantly, it demonstrates the benefits of incorporating more realistic assumptions about the human cognition into a theory of action. In particular, the DPP resonates well with the stated objectives articulated for a sociological theory of action that is more attuned to cognitive aspects. Specifically, the DPP proves to be a profound framework that offers valuable theoretical insights and poses intriguing hypotheses, yet does not require overly sophisticated measures to be meaningfully applied in sociological research. Thus, the DPP represents an exciting opportunity to bridge the gap between complex cognitive understandings and pragmatic sociological applications.

Sammendrag

Denne avhandlingen utforsker kognisjonens rolle i utformingen av menneskelig atferd. Med utgangspunkt i en misnøye med de altfor forenklede antakelsene om menneskelig kognisjon som finnes i mange av dagens sosiologiske teorier, søker denne forskningen å bidra til utviklingen og vurderingen av mer sofistikerte handlingsteorier som bedre fanger opp realitetene i menneskesinnet. Argumentet i denne avhandlingen er at en mer nyansert forståelse av menneskelig kognisjon ikke bare er en intellektuell øvelse, men en uunnværlig forutsetning for å oppnå mer presise og dyptgående sosiologiske forklaringer, samtidig som den gir nye og samfunnsrelevante prediksjoner. For praktisk sosiologisk forskning er det imidlertid avgjørende å sikre at slike berikede teorier forblir håndterbare og ikke drukner i unødvendig kompleksitet. For å nå disse målene presenterer denne avhandlingen seks fagfellevurderte bidrag, knyttet sammen av en innrammende introduksjon, som dekker følgende tre distinkte forskningsområder:

For det første begynner avhandlingen med en kritikk av Rational Choice Theory (RCT), som er valgt som utgangspunkt for denne undersøkelsen fordi den er et typisk eksempel på handlingsteorier som bygger på forenklede antakelser om menneskesinnet. Deretter presenteres to laboratorieeksperimenter som er utformet for å samle inn mer data om menneskelig kognisjon og kan bidra til å forbedre nåværende eller utvikle nye handlingsteorier. Disse eksperimentene fokuserer på menneskets kognitive evne til å resonnerer iterativt, noe som er en kritisk forutsetning for mange av løsningskonseptene som RCT bruker for å forutsi atferd. Studiene viser at menneskets evne til iterativ resonnerement er svært begrenset og at det er liten variasjon mellom individer, noe som står i skarp kontrast til RCTs antakelser, som

vanligvis forutsetter ubegrenset kapasitet på dette området. Som en konsekvens av dette stemmer ikke den observerte atferden overens med forutsigelsene i RCT, noe som understreker viktigheten av å ta hensyn til kognitive begrensninger. Videre indikerer dataene at, i tillegg til kognitive evner, spiller individuell resonneringsstil – det vil si hvordan man benytter sine kognitive ferdigheter – en betydelig rolle i beslutningstakingen. Denne faktoren ser ut til å ha en enda større påvirkning enn selve kognitive evner.

For det andre og som en utvidelse af de foregående innsiktene legger denne avhandlingen økt vekt på studiet av resonneringsstiler. I denne sammenhengen undersøkes to eksisterende teoretiske tilnærminger som tar for seg skillet mellom kognitive resonneringsstiler: Dual Process Perspective (DPP) og Status Characteristics Theory (SCT). DPP er et generelt teoretisk rammeverk som skiller mellom intuitiv og veloverveid resonnering, og som beskriver deres kombinerte effekt på atferd. SCT kan derimot betraktes som en spesialisert anvendelse av DPP, med fokus på hvordan statusindikatorer intuitivt påvirker atferd i fravær av bevisst inngripen. Etter den teoretiske redegjørelsen for begge tilnærmingene avledes spesifikke hypoteser. Angående DPP fokuserer disse hypotesene på betingelsene for at enten intuitiv eller bevisst resonnering kan ha større innflytelse på atferd, og de mulige konsekvensene av en slik innflytelse. SCT brukes til å forutsi hvilke status-signaler som sannsynligvis vil påvirke atferden, og hvilke som ikke vil gjøre det. For å validere disse hypotesene ble det gjennomført ytterligere to laboratorieeksperimenter. Resultatene bekrefter generelt de viktigste prinsippene i både DPP og SCT, noe som understreker verdien av å integrere kognitive aspekter i sosiologiske handlingsmodeller.

For det tredje undersøker avhandlingen om DPP kan brukes i sosiologisk forskning utenfor laboratoriet. Til dette formålet presenterer avhandlingen empiriske bevis fra et ikke-reaktivt felteksperiment og en nettbasert spørreundersøkelse. Felteksperimentet fungerer som en lakmustest for DPPs eksterne validitet ved at prinsippene anvendes i en reell kontekst uten at deltakerne er klar over at de deltar i en studie. Den nettbaserte undersøkelsevalueringen vurderer anvendeligheten av kjernekonseptene i DPP innenfor rammen av standard sosiologiske undersøkelsesteknikker. Begge

studiene bekrefter at DPP er nyttig i virkelige situasjoner og er kompatibel med vanlige nettbaserte undersøkelsesformater.

Konklusjonen er at denne avhandlingen gir tydelige bevis som utfordrer forestillingen om at mennesker har ubegrenset kognitiv kapasitet. Enda viktigere er det at den viser fordelene ved å inkludere mer realistiske antakelser om menneskelig kognisjon i en handlingsteori. DPP stemmer godt overens med de uttalte målene for en sosiologisk handlingsteori som i større grad tar hensyn til kognitive aspekter. DPP viser seg å være et dyptgående rammeverk som gir verdifull teoretisk innsikt og tilbyr spennende hypoteser, men som likevel ikke krever altfor sofistikerte tiltak for å kunne anvendes i sosiologisk forskning. Dermed representerer DPP en spennende mulighet for å bygge bro mellom komplekse kognitive forståelser og pragmatiske sosiologiske anvendelser.

List of Publications

1. Grehl, Sascha, 2020: Verhaltensökonomik und Begrenzte Rationalität (Behavioral economics and bounded rationality). In: Tutić, Andreas (Ed.): Rational Choice. Berlin: De Gruyter Oldenbourg, 150–178.
2. Grehl, Sascha and Andreas Tutić, 2015: Experimental Evidence on Iterated Reasoning in Games. PLoS ONE 10(8): e0136524.
3. Tutić, Andreas and Sascha Grehl, 2017: A Note on Disbelief in Others Regarding Backward Induction. Games 8(3): 33.
4. Tutić, Andreas and Sascha Grehl, 2018: Status Characteristics and the Provision of Public Goods – Experimental Evidence. Sociological Science 5: 1–20.
5. Grehl, Sascha and Andreas Tutić, 2022: Intuition, Reflection, and Prosociality: Evidence from A Field Experiment. PLoS ONE 17(2): e0262476.
6. Tutić, Andreas and Sascha Grehl, 2021: Implizite Einstellungen, explizite Einstellungen und die Affinität zur AfD (Implicit Attitudes, Explicit Attitudes, and the Affinity Towards the AfD). Kölner Zeitschrift für Soziologie und Sozialpsychologie 73: 389–417.

The first publication is reprinted with permission from De Gruyter. All other publications have been published under a Creative Commons Attribution License (CC BY 4.0). All rights reserved.

Contents

1	Introduction	1
1.1	Sociological Explanations	9
1.2	Theories of Action	12
1.3	Quality Criteria	16
2	The Role of Cognition in Sociological Models of Action	21
2.1	The Rational Model of Action and Its Limitations	22
2.1.1	Behavioral Economics and Bounded Rationality	26
2.1.2	Experimental Evidence on Iterated Reasoning in Games . . .	30
2.1.3	A Note on Disbelief in Others Regarding Backward Induction	36
2.1.4	Discussion	40
2.2	Alternative Models of Behavior	41
2.2.1	Dual Process Perspective	43
2.2.2	Dual Process and Prosocial Behavior: Experimental Evidence	49
2.2.3	Status Characteristics Theory	59

2.2.4	Status Characteristics and the Provision of Public Goods: Experimental Evidence	62
2.2.5	Discussion	65
2.3	External Validity and Further Applications	66
2.3.1	Dual Process and Prosocial Behavior in the Field	67
2.3.2	Dual Process and Voting Intentions	69
2.3.3	Discussion	72
3	Conclusion	74
	Bibliography	80
	Articles	105
	Verhaltensökonomik und Begrenzte Rationalität (Behavioral economics and bounded rationality)	105
	Experimental Evidence on Iterated Reasoning in Games	135
	A Note on Disbelief in Others regarding Backward Induction	155
	Status Characteristics and the Provision of Public Goods – Experimental Evidence	163
	Intuition, Reflection, and Prosociality: Evidence from A Field Experi- ment	184
	Implizite Einstellungen, explizite Einstellungen und die Affinität zur AfD (Implicit Attitudes, Explicit Attitudes, and the Affinity Towards the AfD)	199

1 Introduction

Human society is intricate and complex, shaped by the actions of countless individual actors, each with their own unique motivations, backgrounds, abilities, and perceptions of the world. These actions intertwine and converge to form complex patterns, manifesting as social phenomena. In their pursuit to understand and explain these patterns, sociologists employ various theories to uncover the underlying principles that govern these social entities. While some of these theories delve into the micro level, scrutinizing the intricate details of individual interactions, others ascend to the macro level, capturing the broad expanse of societal structures and patterns. Yet despite their different vantage points, nearly all sociological theories share a common foundational element: The integral role of a theory of human agency. This is evident in individualistic traditions such as Behaviorism (Homans 1974) or Rational Choice Theory (Abraham and Voss 2000). It is also discernible, albeit in a more rudimentary form, in approaches such as network theory (Granovetter 1985). Similarly, even macro-sociological paradigms tend to acknowledge the role of individual action. In Parsons' (1937) structural functionalism, for example, the focus is on systems and subsystems and their respective requirements for survival. However, these (sub)systems also provide the normative framework for Parsons' voluntaristic theory of action, in which the behavior of actors is guided by the values and norms of these systems. This serves as a reminder that even within macro-sociological paradigms, individual actions, albeit interpreted through a different lens, are still of profound significance.

Given this pervasive role of action theory, it becomes clear that a comprehensive understanding of social patterns and phenomena, which is the focal point of sociological inquiry (Lindenberg 1992; Wippler and Lindenberg 1987), requires a profound understanding of human action. Consequently, disentangling the nuances inherent in human action is not a secondary task, but a central concern of

sociology. However, the pursuit of understanding human action extends beyond the confines of sociology, permeating numerous neighboring scientific disciplines, such as economics, psychology, or even biology. All of these disciplines share a growing focus in recent years on the cognitive underpinnings of human decision making (e.g., Kahneman 2011; Kenrick and Griskevicius 2013; Rubinstein 2007). Given this interdisciplinary shift, it seems reasonable to use the insights of these disciplines in the pursuit of a more comprehensive and nuanced theory of action. Consequently, this thesis explores the potential of incorporating insights from cognitive research into sociological theories of action with a particular focus on decision making. In doing so, this thesis embarks on a journey that begins with a concern for the cognitive abilities of human decision makers and gradually shifts its focus to an emerging paradigm in research that distinguishes between intuitive and deliberative decision-making behavior.

Yet, the notion of incorporating aspects of human cognition into sociology is not new. Classical sociological theorists have already acknowledged, either explicitly or implicitly, the importance of cognitive aspects in shaping individual perception, judgment, and behavior. Max Weber (1922), for instance, emphasized that people act on the basis of their interpretations of the world around them, thus highlighting the importance of subjective meaning in individual actions and the role of perception in shaping them. Similarly, the Thomas theorem, a seminal concept in sociology, asserts that the subjective interpretations that individuals make of a situation, regardless of their objective accuracy, lead to tangible consequences for the actors involved (Thomas and Thomas 1928, 572).

In a similar vein, Alfred Schütz (1990) sought to understand how individuals' past experiences influence their subsequent behavior. Building on his phenomenological perspective, Schütz (1990, 207ff) posited that individuals construct their reality based on their past experiences and the meanings they attach to them. He further theorized that these past experiences form a repository of knowledge that individuals draw upon when interpreting their current situations and deciding how to act. In his view, therefore, social action is not simply driven by external forces or societal structures, but is also a product of an individual's subjective interpreta-

tion of the world, which in turn is shaped by his or her knowledge gained from past experiences.

Connecting to this, Pierre Bourdieu's (1990) practice theory sheds light on the impact of different types of knowledge on human behavior, highlighting in particular the distinction between mere knowledge and embodied knowledge, as exemplified by the habitus. Bourdieu asserts that internalized cultural knowledge shapes individuals' perceptions as well as interpretations of social situations and guides their decision making. For instance, a working-class individual may be informed about the rules of etiquette of high society, but may lack the practical (embodied) knowledge necessary to implement them in a given social situation due to a lack of socialization.

This account of sociological ideas about the cognitive underpinnings of human behavior is far from complete. There are a plethora of other theories and perspectives in sociology that have implicitly or explicitly acknowledged the significance of cognitive processes in understanding human behavior and social phenomena. Rather than extend this exploration further, it is essential to note that while early sociological work is rich in insights into the potential interplay between cognitive aspects and human decision making, it often falls short in coherently integrating these notions into a comprehensive theory of action. The formulations of these theories sometimes suffer from a lack of clarity and formal rigor, which in turn renders their empirical predictions somewhat ambiguous. This underscores the need for a more systematic and methodical approach to incorporating cognitive aspects into sociological theories of action.

In comparison, Rational Choice Theory (RCT) stands out as a prominent approach within the social sciences, offering a highly developed paradigm for action theory in terms of formal representation (Hedström and Swedberg 1996). At the core of RCT is the decision-making procedure, which we will refer to as *rational decision making*: According to it, individuals have the ability to rank all potential decision alternatives on the basis of a single characteristic, personal preference, and subsequently choose the available alternative that is most preferred (Rubinstein and Osborne 2020). Thereby, RCT rests on a set of axioms that facilitate the precise

formulation and rigorous analysis of human actions and decisions. The successful application of RCT in both theoretical (e.g., Breen and Goldthorpe 1997; Coleman 1990; Diekmann 1985) and empirical (cf. Abraham and Voss 2000; Friedman and Hechter 1988) research impressively demonstrates the usefulness of this approach for advancing sociological knowledge. In this regard, RCT is not limited to a specific area of sociological inquiry, but can be applied to a variety of domains, such as the family (Coleman 1993; Hoem 1991), crime and deviance (Becker 1968; Cornish and Clarke 2014), stratification and mobility (Logan 1996; Walder 1992), or religion (Iannaccone 1991).

At the same time, however, a number of critical contributions to RCT have appeared (e.g., Simon 1957; Smelser 1992). One of the main concerns raised by the critics is the lack of realism in the assumptions underlying the rational decision-making procedure. While these assumptions facilitate the development of elegant and tractable models, they often oversimplify the complexity of human cognition and decision-making behavior. This lack of realism may inadvertently limit the applicability of RCT to understanding real-world social phenomena, thus necessitating a more comprehensive approach that takes into account the intricate nature of human decision making.

In particular, RCT is often criticized for being too individualistic and not taking into account the social context (Granovetter 1985), as well as for assuming that human actors are primarily concerned with their personal material well-being (Münch 2007). Furthermore, it is often argued that RCT assumes that individuals have unrealistically high cognitive abilities (Simon 1955). However, many of these criticisms can be readily dismissed due to misconceptions surrounding the RCT framework. For example, the assumption that human actors are inherently individualistic and materialistically selfish may be common in economics, but it is by no means an inherent requirement of RCT. Instead, the framework can easily accommodate actors who care about the well-being of others (Fehr and Schmidt 1999), immaterial values (Frank 1988), or social context (Buskens and Raub 2013).

With respect to the assumptions of RCT about cognition, however, the critics raise some legitimate concerns. For instance, to make behavioral predictions, RCT em-

employs intricate mathematical models, some of which require considerable computational effort on the part of the researchers. At the same time, however, these same researchers maintain the (at least implicit) assumption that human actors consistently possess the necessary cognitive capacity to solve even the most challenging tasks flawlessly, optimally, and instantaneously (see also Simon 1957).¹ This and similar assumptions are particularly troubling because empirical evidence shows that human actors not merely require more time to make decisions and occasionally make mistakes, but that the inherent cognitive limitations of actors require them to apply entirely different decision-making rules (e.g., Simon 1955). As a consequence, empirical observations of human behavior systematically deviate from the predictions of classical RCT (e.g., Camerer 2003; Thaler 1980).

In response to these criticisms, the Bounded Rationality approach has emerged, positioning itself as a theoretical framework that draws upon the foundational principles of RCT while acknowledging the cognitive limitations inherent in human decision makers (Simon 1990). In particular, the axiomatic branch of this approach (cf. Rubinstein 1998) strives to maintain the advantages of RCT, such as its formal rigor, while providing a more accurate representation of human decision-making processes. Rather than providing a single, universally accepted model, this perspective offers a general framework for developing alternative decision-making procedures that explicitly take into account the cognitive limitations of real human actors (Rubinstein 1998). The Bounded Rationality approach thus provides an appropriate starting point for integrating cognitive insights into a theory of action.

However, it is important that these alternative models do not emerge from an arm-chair perspective that does not relate to reality. Instead, theory building must be based on empirical evidence (Hedström and Swedberg 1996). Consequently, empirical research and experiments are required to gather the necessary data. The

¹Even the staunchest supporters of RCT acknowledge that these assumptions do not accurately reflect reality. Nevertheless, proponents of RCT who adhere to an instrumentalist perspective, such as Friedman (1953), defend these simplifying assumptions. From their perspective, the primary concern is whether a theory can produce compelling and empirically accurate predictions; the degree of realism in the assumptions is considered secondary or even inconsequential. See sections 1.3 and 2.1.1 for a more detailed discussion of this point.

typical approach to gather such data uses RCT to generate predictions for an experimental design derived from the most critical assumptions of RCT, which are then tested for validity (Thaler 1980). When discrepancies are identified, they can serve two purposes: Either to lay the groundwork for formulating alternative assumptions and models (Kahneman and Tversky 1979; Tversky 1972), or to facilitate comparison with pre-existing alternative explanatory models (Bosch-Domènech et al. 2002; Crawford and Iriberry 2007). The first contributions of this thesis follow this principle and focus on the measurement and evaluation of specific cognitive limitations of human actors via laboratory experiments.

Following the initial stage of research for this thesis, however, it became apparent that focusing solely on cognitive constraints provides only a partial understanding of how cognitive factors influence decision making. Both RCT and the Bounded Rationality approach assume that actors always employ (bounded) rational decision making in any given situation. Yet, in our experiments, we observed that while some participants made efforts to obtain sufficient information and took the time for thorough deliberation, others tended to use more intuitive decision-making methods, foregoing critical information or reflection, which resulted in suboptimal outcomes. Thus, it was inferred that besides cognitive capabilities, the style in which actors use their cognitive resources during the decision-making procedure is also of vital significance.

This insight led to a shift in the focus of this research, from the integration of cognitive abilities to the incorporation of reasoning style into a theory of action. Interestingly, similar to considerations of cognition, many ideas about the manifestations of different reasoning styles and their role in everyday human experience can already be found in the seminal works of the social sciences. For example, Vilfredo Pareto (1917) and Max Weber (1922) recognize rational decision making as only one of many types of decision making. Thus, Pareto (1917) distinguishes between logical, non-logical and illogical conduct. Likewise, in Weber's (1922, 24ff) conceptualization, rational decision making, equated with *Zweckrationalität*, constitutes merely one of four ideal types of decision making.

Giddens (1984) not only makes a similar argument, but also qualifies it by positing that *discursive consciousness*, a concept analogous to the rational decision making proposed by RCT, comprises only a minor aspect of human action. Instead, he asserts that the majority of human behavior is based on *practical consciousness*, which involves routinized, quasi-automatic sequences of actions adapted to everyday life. This assertion aligns with Bourdieu (1990), who posits that habitual actions, which are rooted in embodied dispositions and deeply ingrained cultural norms and thus occur unquestioned and quasi-automatically, constitute the bulk of everyday actions.

While numerous renowned theorists have expressed similar ideas in the past, it can be consternated that these ideas usually do not constitute genuine theories of action. In contrast, more recent sociological literature begun to formulate theories of action that explicitly or implicitly incorporate reasoning styles into their frameworks. For example, the model of frame selection proposed by Esser (1996) and further developed by Kroneberg (2014; see also Esser and Kroneberg 2015) offers an explicit approach, suggesting that human actors employ either an impulsive or a reflective reasoning style when making decisions. This model aligns with the *Dual Process Perspective*, a broader paradigm rooted in cognitive and social psychology (Chaiken and Trope 1999; Evans 2010; Stanovich 2011) that has gradually gained traction in economics (Alós-Ferrer and Garagnani 2020; Brocas and Carrillo 2014) and sociology (e.g., Lizardo et al. 2016; Miles et al. 2019; Vaisey 2009) in recent years. By emphasizing the interplay between intuitive and deliberative processes as core components of human reasoning, this perspective provides a comprehensive framework for integrating and interpreting the findings of the research presented in this thesis. An approach in which the distinction between intuitive and deliberative reasoning styles is more implicit can be found in *Status Characteristics Theory* (see Berger et al. 1977). This theory, which can be seen as a specific application of the Dual Process Perspective (Miles et al. 2019), posits that intuitively perceived status characteristics guide actors' behavior unless a deliberative process intervenes (Simpson and Walker 2002; Simpson et al. 2012).

In light of our empirical findings, the subsequent contributions move away from an exclusive reliance on RCT and Bounded Rationality to incorporate the insights of the more specialized Status Characteristics Theory (SCT) and the comprehensive framework offered by the Dual Process Perspective (DPP). The primary objective of these later contributions is to evaluate the potential of these approaches to generate sociologically relevant hypotheses and subsequently to test their empirical validity. As we will see in the remainder of this framing introduction, the DPP in particular emerges as a promising approach for providing the basis for a meaningful sociological theory of action that incorporates cognitive elements.

The central argument put forth in this framing introduction and throughout the thesis is that contemporary sociological theories of action can benefit from systematically incorporating key insights about the cognitive foundations of social action, drawing on both classical sociological action theory and modern cognitive science. To support this claim, the thesis presents six peer-reviewed contributions. The first three contributions focus on addressing the theoretical and empirical limitations of classical RCT, particularly its assumptions about human cognition. Subsequently, the general DPP framework and the specialized SCT are presented and their empirical validity is evaluated through laboratory experiments. The final two contributions build upon the results of the laboratory experiments on the DPP, aiming both to verify the external validity of the laboratory findings and to explore the broader implications and applicability of the DPP approach to survey research.

The subsequent framing introduction to this thesis is structured as follows: First, the remainder of this section provides a conceptual outline of sociological explanations and articulates the necessity of incorporating a theory of action within such explanations. Then, the concept of action theory is explained in more detail and criteria for evaluating its quality are established. In the second section, which constitutes the core of this framing introduction, the individual peer-reviewed contributions of this thesis are summarized and contextualized within the broader framework. In addition, complementary theoretical explanations and empirical

findings are provided. Finally, the last section discusses the findings of this thesis and assesses their implications.

1.1 Sociological Explanations

Sociology is the science that studies how social conditions give rise to social consequences (Lindenberg 1992). Two aspects of this statement are of particular importance: First, the emphasis on *how* signifies that the central focus of sociological analysis should be devoted to uncovering the underlying mechanisms behind social phenomena. This perspective coincides with the principles of explanatory sociology (also called rigorous sociology, cf. Raub et al. 2022), according to which the primary objective of sociological research is to provide causal explanations for social phenomena (cf. Manzo 2021).²

Second, the focus on social conditions and social consequences implies that the analytical primacy of sociology should typically be on social entities and the macro regularities that govern them (cf. Wippler and Lindenberg 1987). This does not mean, however, that sociology is limited to macro processes to explain these macro regularities, since micro processes, especially individual action, are also taken into account (at least implicitly) in many sociological theories. This is because, despite the colloquial tendency to attribute actions to social entities such as groups, organizations, or societies, it is the individuals comprising these entities who act as agents of these entities at the micro level (cf. Zahle 2014).

The debate about the relevance of macro- and micro-level processes is closely related to the ongoing discussion about methodological individualism and method-

²In addition, explanatory sociology rests on the notion of a logical equivalence between explanation and prediction. This means that it is crucial that phenomena are not only explained retrospectively, i.e., after the fact, but that they can also be explained prospectively, i.e., predicted. This aspect underscores the empirical orientation of explanatory sociology, in which explanations, or more generally the theories on which these explanations are based, must be tested against empirical reality (Hedström and Swedberg 1996). Theoretical constructs that are capable of providing explanations but fail to make predictions cannot be tested for validity and, therefore, do not qualify as theories in the sense of explanatory sociology.

ological holism. Methodological individualism stresses the importance of individual processes, while methodological holism emphasizes the role of the macro level. However, it would be an oversimplification to assert that these approaches are diametrically opposed, seeking to explain phenomena exclusively at either the micro or macro level (cf. Udehn 2002; Zahle and Collin 2014).³ In particular, most proponents of methodological holism do not deny the role of micro-level processes in explaining social phenomena (Zahle and Collin 2014, 7). In other words, regardless of the approach taken, micro-processes must be taken into account in sociological explanations. In light of this understanding, it seems appropriate to focus exclusively on micro-level explanatory processes in the subsequent discussion.

Sociological explanations via the micro level are perhaps best illustrated by a schema popularized by James Coleman (1987; 1990) and shown in Figure 1.1. The schema, also known as the micro-macro link (Raub et al. 2011), depicts the logical sequence of explaining (or predicting) a social phenomenon. The upper part of the schema represents the macro level of the social entity being analyzed: On the right, the social consequences to be explained or predicted, and on the left, the preceding social conditions on which the explanation or prediction is based.

According to methodological individualism, these macro consequences cannot be directly explained by macro conditions because social entities do not have agency and cannot causally influence social facts. Therefore, any observed macro regularities (4) are only spurious correlations.⁴ Instead, the explanation must result from the logical connection between the micro and macro levels through the causal processes (1) to (3). These processes have been labeled in the literature with various terms, such as logics (e.g., Esser 1993), assumptions and rules (e.g., Wippler and Lindenberg 1987), or mechanisms (Hedström and Ylikoski 2010). Although these

³While there exist theories that focus exclusively on either the micro or macro level to explain social phenomena, such approaches are uncommon and represent outliers in the discipline (cf. Udehn 2002; Zahle and Collin 2014). One such example is Thomas Hobbes' (1651) theory of the social contract, which posits that social order is generated by atomistic actors without any consideration of the macro level. Conversely, an example that pays no attention to the individual actions of actors is Peter M. Blau's (1977) theory of organization.

⁴Advocates of methodological holism (Zahle and Collin 2014) would disagree with the previous statement and emphasize the importance and necessity of this pathway for a comprehensive explanation.

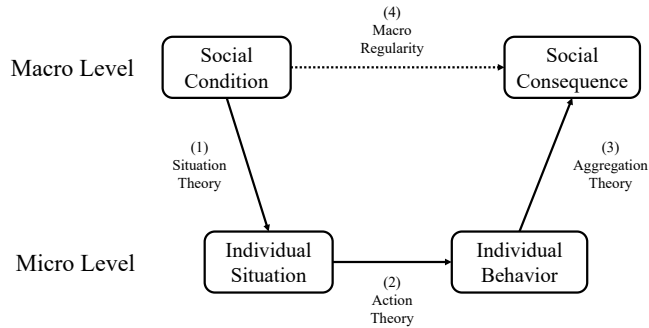


Figure 1.1: Coleman's (1990) scheme of the micro-macro link of explaining macro regularities via a micro foundation.

labels may denote different interpretations of the explanatory processes, for the sake of this analysis we will refer to them as causal processes. In essence, each of the three causal processes must be explained by an appropriate theory.

In the first step of the micro-macro link, the micro conditions for individual actors within a social entity are derived from the macro conditions of the entity. This process requires a theory of situation that enables the identification of features of the individual situation from the broader social condition. In the second step of the micro-macro link, the behavior of actors is to be explained (or, logically equivalent, predicted) on the basis of their individual situations. This requires a theory of action, which will be discussed in more detail in the following section.

Finally, in the third step, an aggregation theory must be applied to explain or predict the macro-level consequences based on the behavior of the individuals. Often, aggregation rules are straightforward and require little explanation, for example, when the collective effect is equal to the sum of independent individual actions, such as the crime rate (Diekmann and Voss 2004). However, aggregation can also be complex due to cascading effects and feedback loops that can arise and affect the macro-level outcome (Coleman 1986; Ylikoski 2021). Various approaches have been proposed to address these challenges (Abell et al. 2008; Manzo 2007; Opp

2011). Although aggregation theory is a crucial component in explaining macro-level outcomes, the framing introduction does not directly address this aspect, as it primarily examines the micro and meso levels.⁵ Note, that this is not meant to downplay the importance of the aggregation process. Rather, it underscores the need to establish a solid foundation in action theory before delving deeper into the complexities of aggregation theory.

Having presented this conceptual outline of the three causal processes involved in sociological explanations through the micro-macro link, we now proceed to explore the second step of this process in more detail.

1.2 Theories of Action

This section explains the concept of action theories in more detail. First, however, we will clarify what we mean by the word *behavior* and other related terms such as *decision*, *choice*, and *action*. While these terms are often used interchangeably in everyday language, they have different meanings in the social sciences. *Behavior* is the broadest term and encompasses any human conduct that originates from the actor himself and is not caused by external forces (cf. Elster 2007, 163f). Thus, climbing a tree is just as much a behavior as is driving a car, but being pulled down by gravity and being moved by the car is not. Note that not every behavior must result in an overt reaction that is visible to an observer. This may be the case if the behavior is aimed at maintaining the status quo, i.e., doing nothing or not changing the current behavior. For example, not coming to the aid of a person in distress is a behavior, even if an observer cannot perceive any external activity.

The term *decision* refers to the process of selecting a specific option from a set of possible behavioral choices, known as the *decision-making* procedure. The option that is ultimately chosen is referred to as a *decision* or, synonymously, a *choice*.

⁵Moreover, certain theories of action are inherently equipped to deal with the aggregation problem (e.g., the Nash equilibrium in game theory, see section 2.1).

Finally, the term *action* in the literature (e.g., Anscombe 1957) usually refers to a subset of behavior that includes only intentional and goal-directed behavior. Elster (2007, 163f), for example, emphasizes the importance of explicitly evaluating different alternatives on the basis of one's own (intentional) goals. Non-intentional or automatic behavior, such as a reflex, is thus dismissed as a phenomenon without scientific relevance for sociology. However, in this framing introduction, which focuses specifically on intuitive behavior, non-intentional behavior is also considered a socially relevant phenomenon. Therefore, we use the terms behavior and action interchangeably.⁶

Having defined the most important terms, we now turn to the theory of action. Figure 1.2 depicts a conceptual framework for a general theory of action that bridges the individual situation and the individual behavior previously discussed in the micro-macro link. While both the situation and the actor's behavior can be directly observed by researchers, the processes that occur within the actor during decision making are not directly observable.⁷ Nevertheless, it is possible to gain insight into these internal processes through the use of mental models, i.e., theories of action. Starting from the individual situation, such models describe the cognitive processes that actors use to perceive the situation (or parts of it), make a choice, and ultimately exhibit a particular behavior.

At the center of the theory of action is what we call the decision-making procedure. It describes how external states (perceived situational conditions) and internal states (dispositional factors) lead to a decision (the choice). In many theories

⁶Moreover, there are several reasons to question the strict dichotomy between behavior and action. First, the threshold for distinguishing between automatic, reflexive behavior from voluntary, unconscious action is not clearly defined (Selten 2002; Sumner and Husain 2008). Second, even proponents of the distinction between intentional and non-intentional behavior do not consistently maintain it, as seen in the interpretation of mixed strategies in game theory as population shares of actors who engage in a particular action without deliberation (e.g., Rosenthal 1979). Finally, from a sociological perspective, the actual consequences of behavior may be more important than the decision-making procedure itself. If automatic behavior plays a significant role in everyday life, as some argue (e.g., Bourdieu 1990), then sociological research should not ignore this type of behavior.

⁷The issue of whether situations and behaviors can be fully and objectively observed is beyond the scope of this framing introduction (for an overview, see Little 1993). Nevertheless, it is reasonable to assume that these external phenomena are easier to observe than the internal processes that occur within individuals.

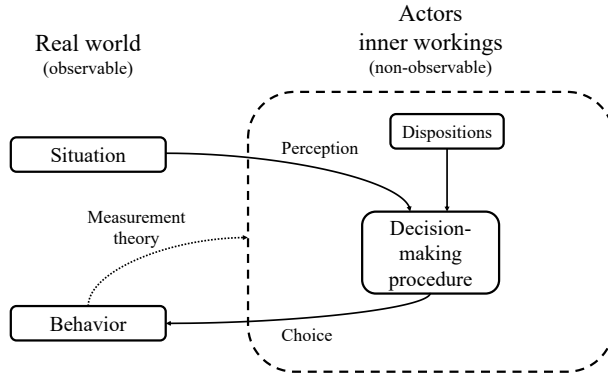


Figure 1.2: Basic framework of a theory of action.

of action, the decision-making procedure is characterized by a sequence of operations that follow law-like rules to derive an action from the internal and external conditions. Individual dispositions, such as preferences, knowledge, and cognitive abilities, tend to be rather dynamic, i.e., they can vary considerably from actor to actor. Nevertheless, action theory can impose certain regularities or constraints on these dispositions.

To help understand this abstract framework, we illustrate it using Rational Choice Theory (RCT). In its simplest form, RCT posits that preferences are the only disposition that holds significance for actors (Rubinstein and Osborne 2020). While actors are relatively free to choose their preferences (Stigler and Becker 1977), RCT imposes certain constraints on their consistency, disallowing some combinations of preferences, such as intransitive preferences. Regarding situational conditions, actors are assumed to fully perceive a situation and derive a set of feasible courses of action from it.⁸ The decision-making procedure is then defined as follows: Actors always choose an action from the feasible set that leads to their most preferred outcome.

⁸In addition to situational conditions, the set of feasible options may also be constrained by dispositional factors, such as the knowledge to perform a particular action.

Let us briefly consider the processes that are also part of the theory of action, namely, the process of perception and the process of choice implementation. Regarding perception, the simplest way to conceptualize it is to assume complete and perfect information, i.e., that actors possess all relevant information. Other conceptualizations may be more realistic, but come at the cost of being more complex and requiring further assumptions.⁹ With regard to the implementation of choice, it should be noted that although we assume a perfect correlation between choice and behavior for the sake of simplicity, this is not necessarily the case in reality. Random factors or hidden motivations may complicate or prevent the inference of choice from the observed behavior.¹⁰

Finally, Figure 1.2 shows a dotted arrow from the observed behavior back to the action theory. This arrow illustrates the idea that an action theory (or any theory in general) must also include a rudimentary theory of measurement in order to have empirical value (Adcock and Collier 2001; Bandalos 2018). A measurement theory allows inferences to be made from the observed behavior of the actor (i.e., the measurement) to all aspects of the model in order to validate them empirically. Measurement theory serves two purposes: First, it defines how the observed behavior can enrich and modify existing knowledge about actors' dispositional factors. For example, new or changed preferences can be inferred from observed behavior. Second, it establishes criteria for determining whether or not an observed behavior is consistent with the theory. This is particularly important because a measurement theory sets the limits of empirical falsifiability and ultimately determines the empirical value of the theory.

Now that we have illustrated the concept of action theory, we can turn to the question of why it is such an important component of sociological explanation. Most obviously, the micro-level transition from individual situation to individual behavior cannot be made without a theory of action. Moreover, the theory of action is also relevant to steps (1) and (3) of the micro-macro link.

⁹We will present an alternative conceptualization in section 2.2.1.

¹⁰Duggan and Levitt (2002) use the case of corruption in sumo wrestling to show that it is indeed very problematic to distinguish the choice "try to win" from "try to lose while trying to look like you want to win" on the basis of directly observable behavior alone.

In particular, the action theory and its decision-making procedure specify which situationally relevant factors may or may not influence the behavior. This, in turn, affects the requirements of a theory of situation in step (1), since it only has to provide information about those aspects that are ultimately taken into account by the theory of action. Thus, if a particular aspect is disregarded by the theory of action, the theory of the situation does not have to consider it. In this sense, action theory preselects the relevant situational aspects. Regarding step (3), some theories of action are already designed to solve the aggregation problem by themselves. For instance, the notion of Nash equilibrium in game theory predicts not only the individual behavior of actors but also states what the collective outcome will be.

Taken together, these points demonstrate the essential role of a theory of action in linking the micro and macro levels of analysis, and thus ultimately in explaining sociological phenomena.

1.3 Quality Criteria

So far, we have outlined the explanatory micro-macro link and emphasized the significance of a theory of action within it. However, the mere recognition of the need for a theory of action does not entail clarity about its essential features or formulation. In this regard, fundamental requirements have been identified that a theory of action must fulfill in order to be considered a theory in the sense of explanatory sociology (cf. Braun 2008). These include, for example, the ability to propose at least one testable hypothesis and the absence of contradictions (Bunge 1996; Popper 1935). Nonetheless, given the boundless possibilities for constructing a theory of action, these rudimentary requirements are insufficient. Therefore, it is imperative to establish objective criteria that can facilitate the scientific selection of feasible models and thus also the scientific progress. In this vein, several criteria for assessing the scientific merit of a theory can be found in the literature that go beyond these basic requirements. In the following we discuss a selection of these

criteria that are particularly important for this framing introduction (for a more comprehensive account: Otte et al. 2023).

The first criterion for evaluating a theory that we will discuss is its *information content*. The information content of a theory is determined by both its *applicability* and the *precision* of the predictions it makes. Applicability refers to the extent to which a theory can be applied across different contexts and situations, which is generally considered desirable because it increases the generalizability of the theory. Specificity of predictions, on the other hand, refers to the degree to which a theory makes precise and narrow predictions rather than vague ones. Narrow predictions are generally preferred in scientific research because they are easier to test and falsify, which is crucial for the advancement of scientific knowledge (Popper 1994).

Related to information content, but not a quality criterion per se, is the degree of *formalization* of a theory. This involves expressing the ideas of the theory in mathematical or procedural terms (such as agent-based models, cf. Manzo 2022). The use of formal models provides a clear and precise language for expressing the concepts, relationships, and predictions of a theory (Olinick 2014), which allows for a more rigorous and systematic analysis of the theory's implications compared to theories that rely solely on verbal arguments.¹¹ In addition, these formal analyses can also facilitate a deeper understanding of the workings of social processes by revealing unforeseen interactions and insights (Aumann 1985).

Another important criterion is empirical *accuracy*, which is a measure of the degree of correspondence between a theory's predictions and empirical reality. This correspondence is often characterized by a continuum that ranges from less to more

¹¹The analysis of the volunteer dilemma illustrates the merits of mathematical models. Regarding the direction of a single effect, e.g., the number of participants on the individual probability of volunteering, verbal theories can provide insight, for example, by using the concept of responsibility diffusion (Darley and Latané 1968) to argue that this effect should be negative. However, when considering the aggregate effect of additional participants on the likelihood of at least one individual volunteering, verbal theories fail to provide direction because it is unclear whether the presence of an additional potential volunteer counteracts the reduced individual probability. In contrast, mathematical analyses (Diekmann 1985; Tutić 2014) can be used to derive precise predictions.

accurate, rather than by a strict dichotomy of right or wrong. A higher degree of accuracy is clearly preferable, since a theory that is consistently contradicted by empirical evidence is not scientifically tenable. Notably, accuracy and precision are interrelated, as less precision cannot lead to less accuracy, and may even lead to more accuracy.¹² Therefore, in scientific research, it is important to strive for both high precision and high accuracy, because both are crucial to the advancement of scientific knowledge.

Let us now focus on two additional scientific criteria, namely parsimony and realism. The *parsimony* or simplicity of a theory is defined as the extent to which a theory contains observable and unobservable constructs (Braun 2008). The smaller the number of such constructs in a theory, the more parsimonious it is. A more parsimonious theory is generally considered preferable to a more complex one, *ceteris paribus*, for a number of reasons. A theory that includes fewer constructs has less restrictive scope conditions, which leads to increased applicability. In contrast, a theory that explicitly assumes the influence of a particular construct may have little or no application to situations in which that construct does not play a role. Moreover, unintroduced constructs do not require measurement, which may be particularly advantageous in situations where it is difficult or impractical to measure such constructs (Lindenberg 1992).

Realism, on the other hand, refers to the degree to which the assumptions, concepts, and relationships formulated in the theory correspond to empirical reality. For example, a theory of action that assumes complete information, i.e., that all actors involved are informed about every relevant detail of the decision situation, is less realistic than a theory that models each actor's individual knowledge.¹³ In addition, realism implies the requirement of causal mechanisms (Hedström 2005; Hedström and Bearman 2009), according to which the processes formulated in the theory should correspond to real-world processes.

¹²To see this, compare the following two statements: "This person behaves prosocially if and only if it is an odd day of the week" and "This person behaves either prosocial or not". While the first statement is quite precise but unlikely to be accurate, the second statement is very imprecise but highly accurate (to the point of tautology).

¹³Note that the assumptions of a theory are often implicit, e.g., a theory that does not consider the concept of social norms implicitly assumes that norms are not relevant.

Scientists generally agree that both parsimony and realism are important scientific criteria for a theory. However, these criteria can be challenging to reconcile because they are often at odds with each other. While parsimony favors simple and concise explanations, realism aims to capture the complexity and nuances of social phenomena. As a result, sociological theories that aspire to be realistic and comprehensive often tend to be more complex in their formulation and presentation, which can undermine their parsimony. Therefore, determining the appropriate weighting of these criteria remains a matter of debate within the social sciences, with instrumentalists and realists at the two opposing ends of the spectrum. Specifically, instrumentalists prioritize parsimony, while realists emphasize realism.

The instrumentalist perspective is supported by the well-known as-if argument put forth by Milton Friedman (1953). In a simplified sense, the as-if argument asserts that it may be reasonable to use an unrealistic theory rather than a more realistic one if the unrealistic theory provides the researcher with a more useful tool for deriving predictions. This argument is based on the notion that the assumptions of a model are ultimately always simplifying and thus empirically incorrect, and that realism is therefore not a useful criterion for evaluating a theory. Since the question of when a theory is realistic enough cannot be answered conclusively, theories should be evaluated primarily according to how many hypotheses (predictions) can be derived from them and the extent to which these hypotheses are empirically validated (cf. Harsanyi 1977). In short, the instrumentalist view holds that the descriptive accuracy of a theory's premises is not crucial, as long as the theory's conclusions accurately predict or match empirical results.

Friedman's (1953) as-if argument has been widely criticized (e.g., Musgrave 1981). Some scholars, such as Hedström and Swedberg (1996), argue that Friedman's argument fails to distinguish between descriptively incomplete statements and those that are descriptively false. While acknowledging that theories will always be incomplete, they contend that theories should strive for descriptive accuracy and not build on descriptively false ideas. In addition, Hedström and Ylikoski (2014, 63) argue that the "primary epistemic goal [of a theory] is to represent the causal pro-

cesses that generate the observable phenomena”.¹⁴ According to this perspective, it is crucial to consider the entire logical chain of an explanation when evaluating a theory. An example of a non-causal explanation is the use of birth control pills to explain why a male person cannot get pregnant. This illustrates the importance of considering the validity of an explanation when evaluating a theory, rather than focusing solely on its ability to make accurate predictions.

In general, the relative importance of the criteria of parsimony and realism, and how to balance them, is not only debated between instrumentalists and realists, but also varies across scientific disciplines and specific applications. For example, psychology, which is primarily concerned with the study of individuals, may prioritize a detailed and realistic theory of human behavior. On the other hand, sociology, which seeks to understand the actions of multiple individuals and their collective consequences, may place a different emphasis on these criteria. With respect to sociological theories of action, Lindenberg (1996) notes that such theories should be designed in such a way that they can be applied to sociological inquiry without requiring the collection of extensive individual-level data. Consequently, a more applicable approach that sacrifices some degree of realism may be preferable in sociology.

In conclusion, the previous elaborations have emphasized the importance of a theory of action in sociological explanation. The criteria for evaluating such theories, including parsimony and realism, have been presented and will guide our assessment of sociological explanations in the following sections.

¹⁴In a similar way, Max Weber (1978, 7) has already pointed out the importance of causal processes: “Sociology [is a] science whose object is to interpret the meaning of social action and thereby give a causal explanation of the way in which the action proceeds and the effects which it produces.”

2 The Role of Cognition in Sociological Models of Action

This section provides a comprehensive overview of the six peer-reviewed contributions that are part of the thesis. Each contribution is briefly introduced and its key aspects are discussed in order to contextualize it within the broader scope of the thesis. In addition, this section provides supplementary arguments and materials that further support the narrative presented in this framing introduction.

The overarching theme of this thesis is to explore the potential of incorporating cognitive aspects into a sociological theory of action and the benefits of such an action theory for sociological research. The material presented can be broadly categorized into three interrelated themes that reflect the evolution of this thesis over time. Each of these themes is discussed in a separate section.

Section 2.1 focuses on a critical assessment of Rational Choice Theory, a prominent theory of action in the social sciences. Following an initial introduction to the fundamental principles of the theory, the section delves into an analysis of the limitations of Rational Choice Theory, particularly emphasizing the constraints arising from its simplifying assumptions concerning cognitive processes inherent in human decision-making. Furthermore, two empirical studies are presented that aim to gain a more detailed understanding of the *cognitive abilities* of real human actors, which may help to identify avenues for the future development of a more comprehensive theory of action.

Based on the findings of these studies, the thesis shifts its focus in section 2.2 from the influence of cognitive abilities to *reasoning style*, a concept that characterizes whether actors are more inclined to reason intuitively or deliberately. This

shift in focus also entails a move away from the previous theoretical framework to theories that have incorporated the aspect of this reasoning style either explicitly or implicitly. To this end, the Dual Process Perspective is presented as a versatile approach alongside the more specialized Status Characteristics Theory. These approaches are briefly introduced and general propositions are derived. Subsequently, their empirical validity is then assessed through two separate laboratory experiments.

Finally, in section 2.3 attention shifts to the broader implications and applicability of the Dual Process Perspective, evaluating its relevance beyond the controlled setting of laboratory experiments. In support of this exploration, two studies are presented. First, a non-reactive field experiment was conducted to establish the external validity of the laboratory findings. This is complemented by an online survey designed to ascertain whether the Dual Process Perspective can be effectively integrated into conventional sociological survey research methods, namely a survey.

2.1 The Rational Model of Action and Its Limitations

The primary objective of the first three contributions of this thesis (Grehl 2020; Grehl and Tutić 2015; Tutić and Grehl 2017) is to provide a critical assessment of the limitations and shortcomings inherent in Rational Choice Theory, a widely used theory of action in the social sciences. Moreover, the aim is to provide insights that can be used in future work to develop an alternative theory of action that can effectively address these limitations.

Before turning to the first contribution, a succinct summary of *Rational Choice Theory* (RCT) is provided.¹⁵ With respect to RCT, we distinguish between parametric *decision theory* and strategic decision theory, which is referred to as *game*

¹⁵A verbal overview will suffice for this summary. For a more formal account, see Rubinstein and Osborne (2020).

theory. In its most basic form, *decision theory* posits that actors consider only their constraints (what they can do) and their preferences (what they want to do) when making decisions. The term constraint includes, for example, available knowledge, time, or money. All these constraints are captured by a single concept, the actor's action space A , which refers to the set of available actions from which an actor can choose. For example, in the context of renting a new apartment, an actor's possibilities are limited not only by situational factors, such as the availability of rental offers, but also by personal restrictions such as her knowledge of where to find such offers, her time to evaluate them, and her financial resources.

Regarding the preferences of actors, it is assumed that actors have a preference relation \succsim over their set of actions A .¹⁶ This preference relation \succsim must satisfy certain consistency criteria, which are expressed in the form of axioms. For example, to ensure that the preferences (and thus the theory) are not contradictory, the preference relation must satisfy the axioms of completeness and transitivity (in which case it is not possible to strictly prefer x over y and at the same time y over x). If the preference relation \succsim satisfies these two axioms and the action space A is non-empty and finite, the preference relation can be represented by a so-called utility function u , a numerical representation of an actor's preferences in which actions are valued higher the more they are preferred.¹⁷ Put simply, this means that actors are able to rank all their actions along a dimension from better to worse.

To behave rationally in a decision situation, as defined by RCT, is to always choose the highest ranked (i.e., the optimal) available action while adhering to the stated consistency criteria for the preference relation. RCT also provides a theory of measurement in the form of criteria that observed behavior must satisfy in order to be explainable by RCT.¹⁸ The purpose of these criteria, as explained

¹⁶A more accurate assumption would be that actors have preferences over the outcomes or consequences of their actions rather than the actions themselves. In decision theory, however, it is often assumed that an action is directly linked to a particular consequence, and thus the intermediate step of preferences over outcomes is often skipped for the sake of simplicity.

¹⁷If the action space A is not finite, \succsim must also be continuous (Rubinstein 2006, 16f) in order to guarantee the representation.

¹⁸The weak axiom of revealed preferences (WARP) is an example of such criteria for the most basic form of decision theory.

earlier in section 1.3, is twofold: First, by observing behavior, new insights (e.g., about preferences) can be gained that can be used to refine predictions. Second, these criteria provide a means of determining whether the observed behavior is rationalizable, i.e., compatible with RCT. Thus, these criteria specify exactly what kinds of empirical observations might falsify RCT.

Note that the two axioms of completeness and transitivity are sufficient to specify an RCT model that can be used to make predictions. However, the information content (see section 1.3) of the model would be very low, since one would not be able to make any predictions other than that people do not make contradictory choices. To increase the information content of the model, additional axioms can be added regarding the preference relation that deal with specific properties of the consequences of actions, such as probabilities or time delays (cf. Kreps and Porteus 1978). The inclusion of such additional axioms not only increases the information content of the model, but also facilitates mathematical operations to be performed on the resulting utility function. This increased mathematical accessibility allows for further analysis and application of the theory.¹⁹

While decision theory (or RCT in general) specifies consistency conditions for preferences, contrary to popular belief, it makes no assumptions about what actors prefer (e.g., Stigler and Becker 1977). Thus, only if actors reveal their preferences through their behavior can certain future actions be ruled out by the model. However, such an agnostic view of actors' preferences is not helpful, especially for sociological analysis (Lindenberg 1996), since one would not be able to make any predictions without first having intensively measured each actor's preferences. Therefore, it is useful to make certain assumptions about the typical preferences of actors, such as the more-is-better assumption (Becker 1971) or the assumption of diminishing marginal utility (Gossen 1854). While the former simply states

¹⁹For example, it does not follow from completeness and transitivity alone that an actor who prefers x over y should also prefer the lottery of getting x with any positive probability and y with the counter-probability over getting y . However, if this is taken as a generally reasonable property of rational behavior, the axioms of expected utility theory (von Neumann and Morgenstern 1944) would lead to such predictions. Moreover, these axioms would allow the ordinal preference relation \succsim to be transformed into an interval-scaled expected utility function u , which in turn would open up additional mathematical possibilities.

that additional quantities of a particular good (in the sense of entities that satisfy human wants and provide utility, such as money, food, or love) are always preferred, the latter states that the increase in utility (i.e., how much it is preferred) decreases as more of the particular good is obtained.²⁰

Game theory is an extension of decision theory that focuses on situations in which the actions of multiple individuals are interdependent and the final outcome is determined by the interactions among them (von Neumann and Morgenstern 1944). Instead of a decision situation, game theory is used to analyze a social situation (called a game) in which a number of actors interact with each other. Each actor i of the set of actors $N = \{1, 2, \dots, n\}$ is characterized, similarly to decision theory, by its own set of constraints A_i and its own utility function u_i for the possible consequences of the game. However, in contrast to decision theory, the consequences of a game are not determined by the actions of one actor alone, but by the actions of all actors involved in the game. Therefore, the preference relation (or utility function) is defined over all possible combinations of actions $A = \prod_{i \in N} A_i$ in this game. A particular consequence (or strategy profile) of a game is then defined by $a = (a_1, a_2, \dots, a_i, \dots, a_n)$, where a_i stands for a particular action chosen by actor i from A_i .

In order to make predictions about the behavior of actors in a game, it is typically insufficient to simply consider the optimal behavior of each actor in isolation, since the optimality of a given action depends on the actions of the other actors involved in the game. To address this problem, game theory uses mathematical solution concepts such as the *Nash equilibrium* (Nash 1950). This solution concept is based on the notion that each actor wants to give a *best response*. An actor i 's best response a_i^* is an action that is optimal given the actor's preferences and expectations about the behavior of others. In a Nash equilibrium $a^* = (a_1^*, a_2^*, \dots, a_n^*)$ each actor's action is a best response to the actions of all other actors. If this holds true, no actor has an incentive to deviate unilaterally from their chosen action, re-

²⁰From a methodological point of view, RCT also requires that preferences exhibit a certain stability, i.e., that they do not change from one moment to the next. Otherwise, any behavioral change could easily be rationalized by (alleged) changes in preferences, which in turn would impair the explanatory power of RCT (cf. Stigler and Becker 1977).

sulting in a stable equilibrium. Note that the Nash equilibrium not only predicts the behavior of individual actors, but also states the collective outcome, serving as both action as well as aggregation theory.

In application, the equilibria of game theory are used as predictions for the individual behavior of actors in a given situation. There are several justifications for using equilibria as predictions. For instance, equilibria are considered more likely to be observed because once actors have reached such a state, they have no incentive to change their actions. Therefore, this behavior is expected to be observable over a longer period of time. In addition, evolutionary dynamics provides reasons for the emergence of equilibria (Maynard Smith 1982). That is, if a strategy profile does not correspond to an equilibrium, it means that there are actors whose actions do not represent a best response. Therefore, they have an incentive to change their actions toward a best response. If all actors follow this reasoning, their actions will eventually converge to a Nash equilibrium, given a sufficient time horizon.

While classical game theory specifies equilibrium states in which each actor's actions represent the best response to the chosen actions of others, the theory does not provide an (explicit) explanation of the process by which actors arrive at this equilibrium state. One possibility is to assume that the actors perform mental simulations of these evolutionary arguments in their minds, and thus all arrive at a choice of action that is part of a Nash equilibrium. However, for the Nash equilibrium to be applicable, especially if we assume that actors perform these mental simulations, several additional assumptions about actors' knowledge and cognitive capabilities are required. These assumptions will be relevant in the next section, where we discuss the limitations of RCT.

2.1.1 Behavioral Economics and Bounded Rationality

The first contribution of this thesis (Grehl 2020) addresses the limitations of Rational Choice Theory (RCT) by identifying several empirical inconsistencies that challenge its basic assumptions. In addition, the contribution reviews a number

of Behavioral Economic models and Bounded Rationality models that attempt to address these inconsistencies. For the purposes of this introductory framework, however, we will focus mainly on the analysis of the shortcomings of RCT supplemented with complementary arguments and a brief introduction to the main ideas of the Bounded Rationality approach. The review of the particular Behavioral Economics and Bounded Rationality models presented in the contribution is omitted, as it is not relevant for the following framing introduction.

In the social sciences, RCT is regarded as a valuable tool because of its ability to formulate precise and empirically testable hypotheses (e.g., Breen and Goldthorpe 1997). While its parsimony allows it to be applied in a wide range of settings (e.g., Becker 1968; Elster 1999; Robinson 1997), its high degree of formalization allows complex relationships and dependencies to be expressed in a concise and precise manner (e.g., Diekmann 1985). On the empirical level, RCT has proven to be extremely fruitful, as evidenced by numerous scientific publications (cf. Abraham and Voss 2000; Binmore 2007) and practical applications of this theory such as auction theory (Milgrom 2004).

Nevertheless, since the emergence of RCT, the theory and its assumptions have been subject to criticism in the social sciences (Simon 1955; Smelser 1992). The criticism includes both direct challenges to the empirical validity of explicitly stated RCT assumptions, such as the transitivity axiom (Tversky 1969) or the axioms of expected utility theory (Allais 1979), as well as challenges to implicitly formulated assumptions. For instance, that actors possess unlimited cognitive abilities that enable them to make decisions without delay, error, or cost (Simon 1976, 1990). Furthermore, it is often implicitly assumed that actors have *complete information* in a given situation, i.e., they are assumed to possess full knowledge of the set of actions and preferences of all actors involved, as well as knowledge of all possible outcomes resulting from the combination of these actions. Moreover, this knowledge is assumed to be *common knowledge* (Aumann 1995; Aumann and Brandenburger 1995) among the actors, i.e., not only does each actor know it, but also everyone knows that everyone knows it, and everyone knows that everyone knows that everyone knows it, and so on ad infinitum.

That the criticism of these assumptions is not merely theoretical is shown by the large number of empirical studies in which a considerable proportion of human actors violate these assumptions (Huber et al. 1982; Loomes et al. 1991; May 1954; Tversky and Kahneman 1981), resulting in a wide range of anomalies (Thaler 1980). Although even RCT proponents know that these assumptions do not correspond to reality, they defend them as well as the resulting anomalies against criticism by arguing, for example, that errors should statistically cancel each other out (Hernes 1992) or that deviations should disappear provided the decision situations are sufficiently simple, the actors sufficiently experienced, and the incentives sufficiently high to motivate RCT-consistent behavior (Binmore 1999). In short, they argue that while the assumptions of RCT may be descriptively inaccurate, they are still close enough to reality to yield coherent explanations and reliable predictions

Indeed, these objections should be taken into account when evaluating anomalies, as there are a number of cases where discrepancies between prediction and observation can disappear under the conditions just mentioned (Goeree and Holt 2001; List 2011). However, not all anomalies can be reconciled in this manner. Evidence exists that a significant portion, and occasionally even a majority, of participants exhibit RCT-deviating behavior that cannot be corrected by experience (Capra et al. 1999; Nagel and Tang 1998), high incentives (Diekmann 2004; Rapoport et al. 2003), or the simplicity of the decision situation (Chou et al. 2009; Costa-Gomes and Crawford 2006). Some deviations are particularly problematic because they are not simple errors that cancel each other out, but rather errors that occur systematically (Thaler 1980; Tversky and Kahneman 1974), meaning that the behavior is biased in a certain direction and can be easily replicated (e.g., Tversky 1969).

An additional counterargument to this criticism is provided by Milton Friedman's (1953) as-if argument, as explained in section 1.3. In the context of RCT, it is used to defend why it may be preferable to model actors as fully informed rational optimizers, even though this may not reflect reality.²¹ Friedman argues

²¹Friedman (1953) also presents an evolutionary rationale in support: Actors who do not behave like rational optimizers would not prevail in the long run. This rationale is often used in

that the demand for an ever more realistic theory of action would make the theory increasingly complex and thus completely useless. This highlights an important objection, but one that can be seen as misguided in the sense that the majority of proponents of an alternative theory of action do not call for more realism for the sake of realism alone. Rather, they argue that the descriptive content of the assumptions should not be seen as an end in itself, but as an indispensable precondition for an improved prognostic and explanatory content of the theory (cf. Opp 1999; Rabin 1998).

In summary, although a parsimonious theory, such as RCT, offers advantages such as increased applicability, it should not be considered sacrosanct because of its simplicity. When empirical evidence supports the validity of a theory in certain domains, but consistently and systematically refutes it in others (see Grehl 2020), it is necessary to critically examine the theory and identify any crucial aspects that have been ignored so far. As a result, it may be necessary to supplement or replace the original theory with a more comprehensive theory that can better explain the observed phenomena and in which RCT is a special case of a more general theory of action. To achieve this, it is essential to identify the aspects that have not been adequately addressed by RCT.

In addition, contribution 1 examines two lines of research that have emerged to overcome the limitations of RCT, namely Behavioral Economics (Camerer 2003) and Bounded Rationality (Rubinstein 1998; Simon 1990). Although the particular models presented in this contribution are not directly relevant to this thesis, a brief examination of the Bounded Rationality approach is warranted, given its significance to the overarching research objectives.

The Bounded Rationality approach strives to preserve the formal rigor of RCT, such as axiomatization, while incorporating other aspects of human decision mak-

the literature to justify RCT assumption. It is based on the assumption that rational optimizers will achieve the best long-term outcomes in any given situation. However, this assumption may not always be valid. While there is some theoretical and empirical evidence to support this rationale in certain contexts (Binmore 2007), there are also instances in which "irrational" behavior, i.e., behavior that does not conform to the axioms of RCT, can be more adaptive (e.g., Bear and Rand 2016; Diekmann 2009; Rand and Nowak 2012).

ing that have been neglected by traditional RCT. In particular, the Bounded Rationality approach attempts to integrate insights from cognitive psychology and related fields to develop models that more accurately reflect the actual decision-making procedure of individuals. In this context, the approach has been concerned with cognitive limitations such as limited perception (Rubinstein 1988), limited memory (Aumann and Sorin 1989), and faulty information processing (Geanakoplos 1992, 2021), which are taken into account in the development of formal models.²²

As such, it provides a promising starting point for the development of more accurate models of decision-making behavior. However, the mere positing of theoretical speculations without empirical support is insufficient. It is crucial to establish theoretical frameworks that are anchored in empirical evidence. In accordance with this perspective, Hedström and Swedberg (1996, 127) assert that “one should always strive to establish a symbiotic relationship between theory and empirical research, where each contributes to the development of the other.” Accordingly, the next two contributions will focus on assessing the cognitive abilities of real human actors in order to contribute to the collective effort to develop more accurate theories of action that better account for the complexity of human cognition and decision making.

2.1.2 Experimental Evidence on Iterated Reasoning in Games

The previous contribution identified limitations in the use of Rational Choice Theory (RCT) as a framework for understanding human behavior and decision making. It argued that a more realistic picture of human reasoning processes may not only be a desirable goal in itself but may also improve the predictive power of a future theory of action that incorporates these insights. Building on this, contribution 2 (Grehl and Tutić 2015) aims to provide a more accurate depiction of human ac-

²²However, some authors such as Reinhard Selten (2002) or Herbert Simon (in Rubinstein 1998, 187ff) do not agree with this interpretation of “Bounded Rationality”.

tors and their decision-making procedure by specifically examining their cognitive capacities.

One such cognitive capacity is the ability to engage in *iterated reasoning* (IR). This ability can manifest itself in a variety of forms, but typically involves actors being able to plan multiple steps in advance and to anticipate the potential actions of others. To do this, the actor must be able to take into account the constraints, preferences, and knowledge of others, and to consider their own actions from the perspective of those others. The ability to engage in IR is often measured by the number of steps involved in the planning or reasoning process. For example, in chess, a player who considers only her immediate advantage is said to be planning one step, while a player who anticipates her opponent's potential countermove and plans a response to that countermove is said to be planning two steps. In principle, this form of IR can go on indefinitely.

The ability to engage in IR is essential for RCT because many game-theoretic solution concepts, such as iterated dominance or backward induction, explicitly require participants to perform at least a certain number of IR steps (Aumann 1995; Binmore 1996). In addition, the Nash equilibrium often implicitly requires IR in the form of common knowledge about various aspects of the game, as emphasized by epistemic game theory (Binmore and Brandenburger 1990; Brandenburger and Dekel 1993; Weber 2001).

The concept of common knowledge is a particularly extreme case, as it requires an unlimited capacity for IR on the part of the actors involved (Aumann 1976). Rubinstein (1989) demonstrates the significant impact on equilibrium predictions that can result when only almost common knowledge, rather than common knowledge, is established. In particular, he shows that even if the process of generating common knowledge is interrupted at a finite but arbitrarily large number of steps due to communication failure, common knowledge cannot be established. Similarly, this argument can also be applied to the ability of IR, where replacing an unlimited ability of IR with an almost unlimited (i.e., finite, but arbitrarily high) capability renders common knowledge impossible, at least in theory. According to RCT, actors in both cases should behave as if common knowledge was not estab-

lished, which in turn can decisively influence the predicted behavior (e.g., Aumann 1998; Diekmann 2009). Thus, obtaining a realistic understanding of the human capacity for IR can not only aid in the development of alternative theories of action but can also help explain deviations from predictions based on unlimited capacities for IR.

Although previous studies have also measured different forms of IR abilities, they have typically assessed only one of these forms at a time (Arad and Rubinstein 2012; Nagel 1995). As a result, they are unable to make statements about potential relationships between different IR abilities. Moreover, some of these studies have not controlled for or even actively encouraged the possibility of learning (e.g., Dufwenberg et al. 2010; Gneezy et al. 2010; Weber 2001), making it difficult to reliably measure the general ability to engage in IR. To address these limitations, in contribution 2 a study was planned that measured the ability to engage in two forms of IR, namely the ability to use backward induction as well as the ability to infer interactive knowledge, while minimizing the possibility of learning. This allowed us to estimate these abilities more accurately and to determine whether or not they are based on a common generalized ability.

In addition, previous studies (e.g., Bosch-Domènech et al. 2002; Brañas-Garza et al. 2012; Ho et al. 1998) that have assessed IR ability have used methods that do not allow for determining whether an individual is exhibiting a particular behavior due to actual limitations in cognitive abilities or due to the belief that their co-actors have reached their cognitive limits and therefore consider it detrimental to allocate further cognitive resources to this type of reasoning. To this end, the measurement protocols were specifically devised to eliminate the influence of varying beliefs about the cognitive abilities of co-actors. This was achieved by using of a fully rational algorithm that served as a substitute for human co-actors in certain parts of the study. Because the participants were informed of this procedure, the cognitive limitations of others could not factor into the measurement of their abilities in these parts of the study.²³

²³Moreover, the public use of rational algorithms as co-players offers several additional advantages: First, all participants can start the interaction under identical conditions, which improves the comparability of the results. Second, other-regarding preferences (e.g., Andreoni 1995b; Fehr

In the laboratory, we measured two forms of IR abilities using two different tasks: For the ability to engage in backward induction the so-called hit game was used (Carpenter et al. 2013; Dufwenberg et al. 2010; Gneezy et al. 2010). In this two-player game, the first player can secure a victory by employing a sufficient number of backward induction steps to derive an optimal strategy. Since participants were always the first player to compete against a rational algorithm, victory was always possible as long as participants adhered to the optimal strategy. Each participant played seven hit games of varying *complexity*, measured by the number of iterative reasoning steps necessary to derive the optimal strategy. As optimal behavior in the game does not depend on any expectations regarding the co-player, it provides an ideal tool for measuring the participants' ability to engage in backward induction.

To measure participants' ability to infer interactive knowledge we used the dirty-faces game (Bayer and Chan Bayer and Chan; Weber 2001). In this game, the participants must infer an initially unknown piece of information based on some initial pieces of information and the behavior of the other participants. Similar to the hit game, this requires participants to engage in multiple steps of IR to infer the information and to solve the game by announcing the correct information.

Unlike the hit game, the dirty-faces game requires participants to take into account the cognitive abilities of others when inferring interactive knowledge. In addition to announcing the inferred information, participants in this game have the option of not solving the game. This option is particularly reasonable when there is suspicion or evidence of inadequate IR ability in others (hereafter disbelief), resulting in the necessary information either not being deducible at all or being too uncertain. Since participants are rewarded for getting the information right, but punished for getting it wrong, it is better for them not to try to solve the game by guessing.

To investigate the impact of disbelief in others on participants' ability to solve the dirty-faces game, two versions of the game were administered to each participant: One with human co-players and one with an algorithm serving as co-players. In and Schmidt 1999) toward a human co-player (e.g., allowing the other person to win) become irrelevant in such a scenario.

the latter case, participants could be confident in the correctness of the algorithm's logical conclusions, so that any failure to solve the game could be attributed solely to a lack of their IR abilities. In the case of human co-players, however, disbelief in others may cause participants to refrain from attempting to solve the game. It is thus expected that a greater number of participants will be unable or unwilling to solve the dirty-faces game in the human version than in the algorithmic version. In each version, participants had to solve seven dirty-faces games of varying complexity.

In addition, we employed the so-called Cognitive Reflection Test (CRT, Frederick 2005), a widely used instrument for assessing individual differences in reasoning styles along an intuitive-deliberative dimension (Kahneman and Frederick 2007). In its original form, the CRT consists of three questions, each of which is designed to elicit an easily accessible and superficially appropriate but incorrect response (see Table 2.1).²⁴ The test requires individuals to engage in some deliberation and critical evaluation of their initial "intuitive" response in order to perform well. Performance on this test is generally interpreted as follows (Rand et al. 2012; Toplak et al. 2011): The higher the CRT score, i.e., the number of correct responses on the CRT, the higher the general tendency to engage in deliberative reasoning. Conversely, a lower the CRT score indicates a higher tendency to use intuitive reasoning.

The results of the study can be summarized as follows: First, we found that individual abilities to engage in IR did not vary much across participants and were more conservative than those found in previous literature (e.g., Levitt et al. 2011; Weber 2001). Specifically, we observed that while most participants were able to reliably engage in one or two steps of IR, only a small minority of 6% of participants could also reliably engage in three steps. Second, participants who performed well on backward induction tasks showed higher performance on problems involving interactive knowledge, suggesting the possibility of a generalized ability for IR. Third, the results indicate that beliefs about co-actors' IR abilities were not relevant in the used dirty-faces game, as the comparison of behavior between the

²⁴In this study, we used the original CRT with three questions. For subsequent studies, we used a modified version with one additional question.

	Question	Correct Answer
CRT1	A bat and a ball cost EUR 1.10 in total. The bat costs EUR 1.00 more than the ball. How much does the ball cost?	5 Cents
CRT2	If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets?	5 Minutes
CRT3	In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake?	47 Days
CRT4	You are participating in a race. You overtake the second person. What position are you in now?	2nd Position

Table 2.1: The original three questions and correct answers of the Cognitive Reflection Test and the fourth question that we added for contribution 3 and subsequent studies.

algorithmic and human versions showed no significant effect. Overall, these results suggest that participants' behavior does not align with traditional RCT and its assumption of unlimited cognitive abilities.

In terms of the influence of reasoning style, we found that participants who scored higher on the CRT, and thus were more likely to engage in deliberative reasoning, performed significantly better in both the hit game and the dirty-faces game. Moreover, individuals with a high CRT score took significantly longer to make the first decision in the hit game. These findings are in line with previous research (Carpenter et al. 2013; Toplak et al. 2011), providing further evidence that reasoning styles play a critical role in the decision-making procedure and thus influence the behavior of the actors. Notably, in the multivariate regression (see contribution 2, Table 10), the CRT score appeared to be more influential for the performance in the dirty-faces game than the IR ability shown in the hit game (i.e., the number of solved hit games). Together with the high variance observed in this measure (see Figure 1a), this result provides initial evidence that the reasoning style of human actors may be at least as important, if not more important, than their cognitive abilities.

As for the null effect of beliefs about the rationality of co-actors, this was somewhat unexpected. Thus, already in this contribution, we hypothesized that the dirty-faces game “is far too complex to allow more or less inexperienced participants to engage in reasoning about the rationality of their co-players” (Grehl and Tutić 2015, 17f). Therefore, we designed a follow-up laboratory experiment to investigate the effect of beliefs in a simpler game.

2.1.3 A Note on Disbelief in Others Regarding Backward Induction

In contribution 3 (Tutić and Grehl 2017), we aim to expand upon the findings from contribution 2 by conducting a more in-depth examination of the influence of beliefs on the iterative reasoning (IR) process. The previous study showed that doubt about the IR capabilities of others (*disbelief*) appeared to have a limited impact on the behavior of our participants. However, it remained unclear whether this result was specific to the game used in that study (the dirty faces game) or whether it was valid in general. Since the existing literature also provides conflicting evidence regarding the extent to which disbelief might affect decision making (e.g., Agranov et al. 2015; Georganas et al. 2015; Levitt et al. 2010; Palacios-Huerta and Volij 2009), another laboratory experiment was planned that focused primarily on the influence of disbelief.

To do this, we modified the hit game we used in contribution 2 to measure the influence of disbelief in the domain of backward induction ability (in contribution 3, this game is called race game). As in the previous study, the participants first played several rounds against a rational algorithm. These rounds were designed in such a way that, in principle, all participants could win each game for sure, provided that they could perform a sufficient number of IR steps to figure out the winning strategy. These initial series of games were used to measure the individual IR abilities.

To assess how certain participants were of winning a given game, they had to choose between two payoff options before each game. Option A guaranteed a fixed prize if the game was won, but also guaranteed nothing if the game was lost. With option B, on the other hand, the participants received the prize with a probability of 70% in case of a win and with a probability of 30% in case of a loss. Thus, option 1 was the better choice as long as the participant believed that winning was more likely than losing.

To examine the influence of the beliefs, the participants then played another series of similar games. Participants were divided into three treatments: In the single treatment participants again played alone against an algorithm. In the team treatment participants were paired into teams, and both team members had to beat the algorithm separately to achieve a win. This means that if either player failed to win against the algorithm, it counted as a loss for both. Lastly, the team-info treatment was similar to the team treatment, but participants were also given the opportunity to inform themselves about the IR ability of their team member.²⁵ Since this procedure was known to all, participants in the team and team-info treatments had to consider not only their own IR ability but also that of their human team member when choosing their payoff option. If, for example, a participant was confident that she would beat the algorithm but her team member would lose, payoff option B was more beneficial.

By examining the differences in the chosen payoff options across the treatments, it is possible to infer the relative importance of disbelief in this game. If disbelief in others has a genuine impact on an individual's behavior, then we expect the following consequences for the second series of hit games: First, in both the team and team-information treatments, option B should be chosen more often than in the individual treatment. Second, in the team-info treatment, observing a low IR ability of the team member should increase the likelihood of choosing option B relative to a high IR ability.

²⁵The information presented was the games won by the team member in the first series. To test whether participants cared about this information, it was initially hidden until they manually revealed it.

In line with contribution 2, our results show that participants exhibited very limited IR abilities, and that these abilities did not differ greatly between participants. This again, refuted the standard RCT assumption of unlimited cognitive abilities. Moreover, we found strong evidence for disbelief in others, i.e., participants in the team and the team-info treatments were significantly more likely to choose option B than participants in the single treatment. This effect was more pronounced in the team-info treatment than in the team treatment. Regarding the influence of information, participants who received information adjusted their behavior accordingly. More specifically, the lower the team member's ability to think iteratively, the more likely participants were to choose option B. Interestingly, the study also found that the participants generally overestimated their team member's ability, i.e., participants who had no information about their team member's ability behaved identically to participants who knew they were teamed up with a relatively skilled team member. In addition, we found that the more complex the hit game, the smaller the effect of disbelief. This observation is reasonable: As the complexity of the game increases, participants' confidence in their ability to win against the algorithm decreases, thereby reducing the importance of their disbelief about their team member's competence.

These results indicate that participants consider the cognitive abilities of others when engaging in problems requiring backward induction. As such, this study adds to the ongoing discourse on the influence of disbelief in others on out-of-equilibrium behavior, emphasizing the need to consider such factors into account when predicting or explaining human behavior in strategic contexts. In light of the existing literature, the study offers two potential explanations for the contradictory findings on the impact of disbelief in others: First, our results show that the effect of disbelief diminishes as problem complexity increases, which could account for varying findings across studies using problems of differing complexity.²⁶ Second, it is possible that the observed overestimation of others' abilities also plays a role and may vary depending on study designs, although more research is needed to explore this possibility further.

²⁶This supports the idea that the dirty-faces game used in contribution 2 was too complex for disbelief in others to matter.

<i>Number of solved hit games</i>	Model			
	(1)	(2)	(3)	(4)
CRT score	0.661*** (0.111)			0.560*** (0.113)
Digit span test result		0.193+ (0.100)		0.070 (0.097)
Mini-q test result			0.049** (0.019)	0.028 (0.019)
R^2	0.159	0.019	0.035	0.175

Notes: $N = 188$, OLS coefficients, standard errors in parentheses

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 2.2: Relationship of cognitive measures and the number of solved hit games in contribution 2 (Tutić and Grehl 2017).

Similar to contribution 2, but not reported in the publication, we also employed the Cognitive Reflection Test (CRT) to assess participants' reasoning style (intuitive vs. deliberative). Furthermore, we measured two additional cognitive abilities, namely short-term memory capacity using a digit span test (Devetag and Warglien 2003) and speeded reasoning ability via the mini-q test (Baddeley 1968; Baudson and Preckel 2016). Table 2.2 presents three bivariate and one multivariate regressions analyzing the effect of reasoning style and ability on performance in the hit game (number of games solved could range from 0 to 14). Consistent with previous studies, the bivariate analyses reveal a positive influence of deliberative cognitive style and cognitive ability on IR performance (Carpenter et al. 2013; Devetag and Warglien 2003; Toplak et al. 2014). However, in the multivariate model, only reasoning style remains a significant factor. Further analyses of reasoning style revealed a significant correlation between CRT score and response time before the first move. Participants with a maximum CRT score took on average 43 longer than those with a minimum score. These findings again suggest that an individual's willingness to engage in deliberative reasoning may have a more significant impact on their performance than their cognitive ability alone.

2.1.4 Discussion

In this section, we have provided a short introduction to Rational Choice Theory (RCT) and presented three peer-reviewed contributions. The first contributions emphasized that traditional RCT, in its desire for parsimony, relies on oversimplified assumptions about the cognitive abilities and processes of real human actors. While we agreed with the proponents of RCT that the pursuit of a more precise theory of action should not be an end in itself, especially if the predictions are consistent with empirical evidence except for occasional random errors, we also emphasized that advocates of a more realistic theory of action seek it primarily because RCT manifests systematic (rather than merely random) deviations that imply an essential lack of critical components in the theory of action. One such component, namely the cognitive abilities of actors, was identified as a worthwhile area for further investigation.

The issue of limited iterated reasoning abilities illustrates the influence of cognitive abilities on human behavior, as demonstrated by the two experimental contributions 2 and 3. In both articles, a laboratory experiment was conducted to obtain a realistic picture of the cognitive abilities of human actors, specifically with respect to iterative reasoning. The results showed that actors could reliably perform only a very limited number of iterated reasoning steps and that there was little variation in the ability of the participants. These findings may explain why equilibrium predictions that rely on high or unlimited abilities to engage in iterative reasoning, such as backward induction, have fared poorly in experimental studies (cf. Fey et al. 1996; McKelvey and Palfrey 1992).

In terms of individual differences in behavior, the findings suggest that cognitive abilities for iterated reasoning may have little explanatory power when all actors have similarly limited abilities. In other words, the extent to which individuals differ in their iterated reasoning abilities may not be significant enough to account for differences in behavior. This result could have implications for the development of action models and the design of experimental studies. While it may be useful

to include a general level of iterative reasoning ability in an action theory, it may not be necessary to incorporate individual differences in this ability.

In contrast, our research also showed that the way individuals choose to employ their cognitive abilities is at least as important as the abilities themselves. That is, the two experimental contributions identified a notable relationship between an individual's tendency toward either an intuitive or a deliberative reasoning style and their behavioral tendencies. Specifically, our results indicated that participants with a high score on the Cognitive Reflection Test (CRT), indicating a deliberative reasoning style, had longer response times on average and demonstrated a greater ability to engage in iterative reasoning. Moreover, our research showed that the CRT score had a greater influence on rational decision making in our tasks than did cognitive ability. This suggests that the way in which participants chose to employ their limited cognitive abilities had a greater impact than the abilities themselves.²⁷

These observations served as a starting point for further investigations into the nuanced mechanisms underlying intuitive and deliberative decision-making processes, which are explored in more detail in the next section.

2.2 Alternative Models of Behavior

The previous contributions have shown that human actors display limited cognitive abilities when it comes to iterative reasoning, and that these abilities exhibit only little variance. While these observations provide fundamental insights for the development of an alternative theory of action that is based on more realistic assumptions about human cognition, they also suggest that individual differences

²⁷However, it should be clarified that the emphasis on reasoning style does not diminish the importance of cognitive abilities in general. First, cognitive abilities are a necessary prerequisite for performing well in our experiments. Yet, they are not a sufficient condition on their own, as actors must also demonstrate a willingness to use these abilities, i.e., to apply a deliberative reasoning style. Second, the scope of our experiments is limited to a particular set of cognitive abilities, specifically the ability to reason iteratively.

in iterated reasoning abilities may not be that important in terms of individual behavioral differences. In contrast, the Cognitive Reflection Test (CRT), used to measure participants' tendency toward intuitive or deliberative reasoning, showed more variance and was more important for the successful completion of iterated reasoning tasks than participants' cognitive abilities themselves. Consequently, these findings prompted a subsequent shift in the focus of this thesis from cognitive ability to reasoning style.

Although the Bounded Rationality approach attempts to account for cognitive limitations through the extension of Rational Choice Theory (cf. Rubinstein 1998), it lacks a similar tradition when it comes to reasoning styles.²⁸ Therefore, in this section, we explore alternative theoretical approaches that either implicitly or explicitly incorporate intuitive and deliberative choices in decision making. Specifically, we examine the general framework of the *Dual Process Perspective* (e.g., Chaiken and Trope 1999; Evans and Frankish 2009; Smith and DeCoster 2000) and the more specific *Status Characteristics Theory* (cf. Berger et al. 1977; Simpson et al. 2012).

Both approaches are used to investigate prosocial behavior in small groups. *Prosocial behavior* is a multifaceted concept that has been studied across various scientific fields, including sociology, psychology, economics, and biology (cf. Padilla-Walker and Carlo 2014). As a result, conflicting definitions and perspectives on the concept exist in the literature (e.g., Batson and Powell 2003; Bierhoff 2008; Levine 2012). In the context of this framing introduction, prosocial behavior is conceptualized as an behavior that meets two necessary conditions: First, the behavior must provide a benefit to another individual or group. Second, the actor must forgo an immediate material or immaterial gain that could have been obtained by choosing a different behavior. The second condition is important because it prevents purely selfish behavior that happen to benefit other people from being considered prosocial.

²⁸Despite some attempts by authors such as Tutić (2015a) to address this issue, the field is still nascent and requires further development.

In order to assess the empirical validity of the two theoretical frameworks, and to evaluate their potential to enhance our sociological understanding, a separate laboratory experiment was conducted for each approach. The central aim was to investigate the potential of these frameworks to offer supplemental explanatory insight into prosocial behavior. Subsequently, a succinct introduction to each approach is provided, accompanied by a presentation of the respective laboratory experiment.

2.2.1 Dual Process Perspective

In this section, we discuss the *Dual Process Perspective* (DPP), a theoretical framework that describes the human mind as governed by two distinct types of cognitive processes. We prefer the term “perspective” over “theory” to acknowledge that no single version of this approach can claim sole authority within the literature (cf. Chaiken and Trope 1999; Lizardo et al. 2016). The origins of the DPP can be traced back to cognitive and social psychology (Chaiken and Trope 1999; Evans 2010; Stanovich 2011), but the DPP has recently also gained traction in sociology (e.g., Murray et al. 2011), particularly in the field of culture (e.g., DiMaggio 1997; Lizardo et al. 2016; Vaisey 2009), and in other scientific fields (e.g., Brocas and Carrillo 2014; Grayot 2020). The fundamental premise of the DPP is that the human mind operates using two distinct types of cognitive processes, which are distinguished by their unique characteristics and functions. Following Evans and Stanovich (2013), we refer to them as *Type 1* and *Type 2* processes.²⁹

Type 1 processes are typically characterized by being fast, automatic, or effortless, whereas Type 2 processes are slow, controlled, or effortful (e.g., Kahneman 2003, 2011). As these characterizations are not without controversy (cf. Keren and Schul 2009; Kruglanski and Gigerenzer 2011; Osman 2004), Evans and Frankish (2009)

²⁹Other labels for these two types of processes are *System 1* and *System 2* (Kahneman 2011; Stanovich and West 2000), *automatic-spontaneous* vs. *reflecting-calculating* (Esser 1996, 2010; Kroneberg 2005, 2011), or *instinctive* vs. *contemplative* processes (Rubinstein 2007, 2013, 2016). Moreover, it is important to note that we intentionally use processes in the plural, as we agree with authors such as Evans and Stanovich (2013), who suggest that human reasoning is the product of an intricate interplay of a variety of cognitive processes.

suggest that these descriptions should be considered as typical correlates rather than invariant properties of these types of processes (see also Evans and Stanovich 2013). Nevertheless, according to Evans and Stanovich (2013), there are crucial defining features that distinguish Type 1 from Type 2 processes. According to them, Type 1 processes run autonomously without the individual being able to inhibit them and without the need to use working memory. In contrast, Type 2 processes must be used intentionally and require the use of working memory.³⁰

The dichotomy between Type 1 and Type 2 processes can be applied to various cognitive domains, including learning, remembering, reasoning, and decision making (cf. Lizardo et al. 2016). For example, the human long-term memory systems, which are critical for all of the aforementioned domains, can be distinguished into declarative (explicit) and nondeclarative (implicit) memory (Squire 2004; Squire and Wixted 2011). *Declarative memory* stores knowledge that requires conscious and controlled retrieval, such as facts and the temporal order of events. In contrast, *nondeclarative memory* is acquired and used unconsciously and stores knowledge such as skills, habits, or dispositions. It is more accessible to Type 1 processes than to Type 2 processes (Lizardo et al. 2016), or as Vila-Henninger (2015, 244) formulates it, “expressed through action rather than conscious recollection”. Both types of memory play an distinct role in the model of action that we will elaborate in the following.

In the context of decision making, DPP posits that decision making involves both deliberative Type 2 processes as well as more intuitive Type 1 processes. The literature, however, one can find divergent views on how these two types of processes affect decision making in the human mind (e.g., Gilbert 1999; Smith and DeCoster 2000). In this framing introduction, we use a model that is based on the well-established default-interventionist model of cognition (Evans and Stanovich 2013) to illustrate how cognitive processes interact. The model (see Figure 2.1) asserts that the decision-making procedure always commences with Type 1 pro-

³⁰To see the difference between correlates and defining features, consider the task of calculating 73×37 . While most people will perform this task relatively slowly, some can do it quickly due to practice and appropriate techniques. In both cases, however, Type 2 processes requiring conscious activation are necessary to complete the task.

cesses that automatically evaluate the situation and intuitively suggest a course of action, referred to as the *default*. Whether this default is ultimately implemented, however, depends on whether Type 2 processes intervene and, if so, whether they maintain, modify, or override the proposed default. In order to describe this model accurately, three specifications are necessary: First, which defaults are typically selected by Type 1 processes; second, the consequences when Type 2 processes intervene; and third, the conditions under which these processes intervene.

As noted above, Type 1 processes are characterized by their automaticity and rapid processing speed, which in turn affect their operation, particularly with respect to situational perception or memory access. In terms of perception, Type 1 processes, due to their speed, usually do not allow for a complete perception of all relevant aspects of a situation, but must focus on the most salient aspects (Strack and Deutsch 2004). However, the degree of salience of different aspects depends on both situational as well as dispositional factors (Fazio 1990). Situational factors, such as framing (see Goffman 1974) and priming (see Bargh 2006), influence the salience of certain aspects of the situation and thus affect behavior (e.g., Liberman et al. 2004; Schwerter and Zimmermann 2020). Dispositional factors, such as optimistic attitudes, may increase the salience of potential benefits of an action, whereas pessimistic attitudes may increase the salience of potential losses. Regarding memory, Type 1 processes are thought to rely more on nondeclarative memory than on declarative memory (Smith and DeCoster 2000) because, as noted above, the latter is not readily accessible without available working memory (cf. Evans 2009). Thus, knowledge stored in the nondeclarative memory (hereafter, nondeclarative knowledge) should influence Type 1 processes more strongly than knowledge stored in the declarative memory.

Type 2 processes, on the other hand, are characterized by a conscious mode of decision making. Unlike Type 1 processes, which are automatic, Type 2 processes can be extended for as long as necessary, allowing for a more thorough assessment of a situation, beyond its most salient aspects. Consequently, Type 2 processes should be less susceptible to the effects of framing and priming, which influence the salience of certain aspects. Furthermore, Type 2 processes are capable of

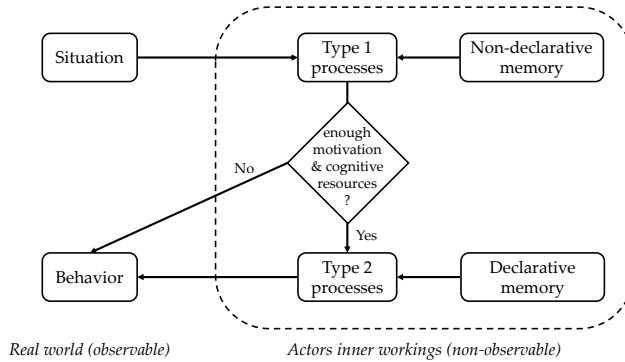


Figure 2.1: A dual process model of action based on Evans’s (2011) default-interventionist model.

integrating knowledge from the declarative memory (Evans 2009), reducing the importance of nondeclarative memories in contrast to the default, and allowing for the evaluation of conflicting information and the potential consequences of one’s actions.

With respect to the potential intervention of Type 2 processes, the DPP argues that deliberative Type 2 processes are costly in terms of cognitive effort and time, which makes people generally inclined to use them as infrequently as possible.³¹ However, the DPP also acknowledges that Type 2 processes can intervene under specific circumstances, such as in novel situations or when the stakes are high. Factors such as personality type, cognitive capacity, and available time may also influence the likelihood of Type 2 interventions (Evans 2011). Empirical research shows that Type 2 processes are less likely to intervene when an individual’s cognitive resources are depleted (Greene et al. 2008), otherwise engaged (De Neys 2006), or when individuals do not have sufficient time for reflection (Rand et al. 2012).

³¹This consideration also plays a significant role in providing a rationale for the evolutionary advantage of human actors developing two distinct types of processes, rather than solely relying on deliberative processes. Given the assumption that deliberative decision making, in contrast to intuitive approaches, entails higher costs in terms of time, energy, etc., it can be hypothesized that the presence of dual decision mechanisms that are used based on situational conditions may constitute an evolutionarily stable strategy (cf. Bear and Rand 2016).

As a result, the decision-making procedure (and the resulting action) can be characterized as either: a) *purely intuitive*, meaning that only Type 1 processes are involved without interference from Type 2 processes; b) *purely deliberative*, with Type 2 processes completely overriding the Type 1 default; or c) *mixed*, meaning that the default suggested by Type 1 processes is modified by Type 2 processes without being completely overridden.³² Note, that this characterization is clearly an idealized abstraction. Evans and Stanovich (2013), for example, emphasize that both types of processes are always involved. For the purposes of hypothesis generation, however, this simplified representation serves as a valuable heuristic. Based on these elaborations, the following proposition can be derived (see also Tutić 2022).³³

Principle of Catalyzation Situationally salient aspects as well as nondeclarative knowledge have a greater influence on behavior when decision making is purely intuitive than when it is purely deliberative or mixed.

This principle thus provides the basis for the relationship between the multiple influences that shape human behavior, emphasizing the interplay between situational elements, knowledge, and decision-making processes. It identifies the conditions that lead to increased reliance on salient aspects of the situation and nondeclarative knowledge. At the same time, it suggests that a shift to more deliberative or mixed decision making patterns reduces the weight of these factors.

Prior to proceeding to an empirical validation of the DPP in the next section, it is worth showing how some canonical ideas in the sociological literature are reflected in the theoretical framework of the DPP. Scholars such as Lizardo (2004) and Vaisey (2009) identify a broad alignment between the DPP and Pierre Bourdieu's

³²An example of a mixed decision is the so-called anchoring phenomenon, where an action is influenced by previously presented but unrelated information (Furnham and Boo 2011). English et al. (2006), for instance, report in the context of judicial decisions that experts' sentencing decisions are influenced by irrelevant sentencing demands, even when they are apparently randomly determined, as in the case of dice thrown by the experts.

³³Note that this proposition cannot necessarily be derived from every DPP model, but only from those that are compatible with the default interventionist model (e.g., Esser 1996; Fazio 1990; Kroneberg 2005).

(1984) practice theory, such as the dispositional nature of perception and action. Thus, Type 1 decision making based on nondeclarative knowledge can be linked to Bourdieu's (1984) notion of *habitus*, as the habitus represents fast and automatic cognitive processes arising from embodied knowledge (cf. Vaisey and Lizardo 2010). Similarly, Giddens' (1984) idea of *practical consciousness* or Schütz's (1944) idea of behavioral recipes that are based on the *thinking as usual* can be seen as a variant of Type 1 decision making.

As noted earlier, Erving Goffman's (1974) concept of *framing* can be linked to the notion of how salient aspects of a given situation impact Type 1 processing. Moreover, the postulate that nondeclarative knowledge formed through cumulative past experience can influence perception, judgment, and ultimately behavior via Type 1 processes, shares notable similarities with the tenets of phenomenological sociology and symbolic interactionism. In particular, Alfred Schütz (1990) detailed how an individuals' *stock of knowledge*, amassed through past experiences, serves as the foundation for interpreting and engaging with the present social reality. Similarly, symbolic interactionism, as propounded by Herbert Blumer (1986; see also Mead 1934), emphasizes the pivotal role of past experiences and social interactions in shaping individuals' subjective understanding of the world and their actions within it.

With regard to Type 2 processes, two main lines of interpretation can be identified. First, some scholars (cf. Vaisey 2009) draw parallels between Type 2 processes and Ann Swidler's (1986) concept known as the *tool-kit* approach. According to this view, Type 2 processes serve a post-hoc rationalization function rather than actual decision making. Essentially, they are seen as mechanisms through which individuals rationalize and legitimize the choices and actions that have already been selected through intuitive Type 1 processes. This echoes the idea of *discursive consciousness* which is also found in Anthony Giddens' work (1984). Here, Type 2 processes (discursive consciousness) are not necessarily seen as the drivers of decision-making but rather as the means through which people construct narratives and justifications for decisions made at a more subconscious level by Type 1 processes (practical consciousness).

The second interpretation of Type 2 processes, on the other hand, assumes a deliberative and evaluative decision-making process as described in Rational Choice Theory (RCT). In this view, Type 2 decision making involves individuals engaging in a thorough evaluation of possible actions and carefully weighing their potential outcomes before making a choice, which closely mirrors the core tenets of RCT. In fact, some scholars, such as Esser and Kroneberg (2015), view RCT as a special case of pure Type 2 decision making. Our understanding and modeling of the DPP are more consistent with this interpretation.

It is noteworthy that the DPP can not only incorporate these different sociological ideas associated with Type 1 and Type 2 processes into a single framework, but also specify the conditions (such as environmental influences or individual characteristics) under which either intuitive Type 1 or deliberative Type 2 decision making is more likely to occur. In this way, it is a rather universal approach that has the potential to reconcile these two distinct facets of human decision making in an analytical framework, amenable to rigorous mathematical scrutiny (e.g., Esser and Kroneberg 2015).

Having discussed the theoretical underpinnings of the DPP and its potential for integrating various ideas from different fields of sociology, it is now important to test the validity of the framework empirically. In the following section, we will derive hypotheses based on the principle of catalyzation and test them by means of a laboratory experiment, in order to gain a better understanding of the extent to which the DPP can explain social behavior and decision making, and its potential implications for the field of sociology.

2.2.2 Dual Process and Prosocial Behavior: Experimental Evidence

In this section, we present a laboratory experiment designed to test the principle of catalyzation in the context of prosocial behavior. According to this principle, situational focal aspects as well as nondeclarative knowledge should have a greater

impact on behavior when decision making occurs via Type 1 rather than Type 2 processes. As this study is not included in the attached contributions, we will provide a brief description of the experiment by focusing only on those aspects that are relevant to this framing introduction.³⁴

To measure prosocial behavior, we used a one-shot public goods game (PGG, Andreoni 1995a; Olson 1965). This game involves the formation of randomly assigned groups of n individuals, who anonymously, simultaneously, and independently decide how much of their initial monetary endowment to allocate to a public fund. The sum of the allocated money to the public fund is then multiplied by a factor k (with $1 < k < n$) and distributed equally among all group members, regardless of their individual contributions. As a result, each participant receives $\frac{k}{n}$ units of money for each monetary unit allocated to the public fund. Thus, if all group members contribute their entire endowment to the public fund, everyone will receive k times their initial endowment. This situation is clearly better for everyone than the situation in which no one contributes and therefore everyone receives only the initial endowment. However, since $\frac{k}{n}$ is less than 1, there is a cost associated with contributing. Hence, all group members are individually better off if they contribute nothing.³⁵ Therefore, the size of the share of the initial endowment contributed to the public fund can be interpreted as a measure of prosocial behavior.

To experimentally vary the dominant type of cognitive processes involved in decision making, we instructed participants either to choose their allocation as quickly as possible (*pressure condition*) or to think thoroughly before deciding (*leisure condition*). As a point of reference, we told participants in the pressure condition to decide within 10 seconds whereas participants in the leisure condition should contemplate for at least 10 seconds before deciding, following the procedure used by Rand et al. (2012). The primary objective of this manipulation was to facili-

³⁴This manuscript is currently being prepared for publication.

³⁵Note that if we assume rationality and materialistic egoism, the game-theoretic prediction, i.e., the Nash equilibrium, suggests that all group members keep everything for themselves and everyone ends up with only their initial endowment. However, this Nash equilibrium is not Pareto-optimal (Diekmann 1992).

tate intuitive Type 1 processes in the pressure condition and deliberative Type 2 processes in the leisure condition.

According to the principle of catalyzation, both nondeclarative knowledge as well as salient situational aspects are expected to exert a stronger influence on behavior in the pressure condition than in the leisure condition. With respect to nondeclarative knowledge, our focus was on a particular cultural orientation stored in the nondeclarative memory, namely *prosocial attitudes*. Attitudes are typically defined as positive or negative evaluations of an object or category of objects (Eagly and Chaiken 1993; Fazio 1995). In sociology, attitudes toward social categories such as gender and ethnicity or toward social concepts such as prosocial behavior or racism are of particular interest. In this study, we measured participants' attitudes toward prosocial behavior via a 13-item version of the prosocial personality battery (Penner et al. 1995).³⁶ We then calculated a Prosocial Attitude Score (PSA score) for each participant, which could range from 0 (minimally prosocial) to 1 (maximally prosocial).

Regarding situationally salient aspects, we simultaneously varied the priming of participants prior to the PGG as well as the framing of the PGG itself. In the *neutral condition*, both the priming stimulus and the framing of the PGG instruction were designed to be as neutral as possible. In contrast, in the *cooperative condition*, the priming and the framing were designed to make prosocial behavior more salient.

Participants were primed with the scrambled sentences task (SST) immediately prior to playing the PGG. The SST is a commonly used priming technique (cf. Shariff and Norenzayan 2007; Srull and Wyer 1979) designed to subconsciously make participants more receptive to certain aspects and thus influence their response in a subsequent task – in our case the PGG. Participants were presented with five words in a scrambled order and were asked to form a grammatically

³⁶See S3 Table in contribution 5 for more details. To control for potential order effects, we measured the attitudes of approximately half of the participants via an online questionnaire at least four days before the laboratory experiment, while the other half had their attitudes measured immediately after the experiment. We found no evidence of differential effects based on measurement timing.

correct four-word sentence with these words. In the neutral condition, the SST consisted of 15 sentences that were as general as possible and therefore unlikely to influence participants' decisions in the PGG in any coherent way. In the cooperative condition, ten of these sentences were replaced with sentences such as "together everything is easier" or "honesty lasts the longest", which primed a prosocial sentiment and thus made a prosocial response in the PGG more salient (e.g., Abbate et al. 2013).

In our study, we manipulated the framing of the PGG by altering the instructions given to participants, following the approach of Liberman et al. (2004). Specifically, we made three key changes across framing conditions: First, the name of the game was either "investment game" or "team game"; second, the group of three was either referred to as "set of players" or "team"; and finally, contributing to the public fund was either referred to as "investing" or "contributing to a team project" in the neutral and the cooperative conditions, respectively.

In addition, we also varied the incentives in the PGG by fixing k at 1.2 (low incentive condition) or 2.1 (high incentive condition). Since these incentive conditions are not of particular interest for this framing introduction, we use them only as control variables.

At the end of the experiment, participants completed a questionnaire that included an extended version of the Cognitive Reflection Test (Frederick 2005) to assess their general disposition toward intuitive versus deliberative decision making. This test included the original three questions plus an additional question (see Table 2.1). Although this measure did not play a role in the current experiment, the fact that we measured it in this experiment will be relevant to contribution 5.³⁷

Having described the experimental setup and design, we now establish a connection between the experiment's features and the DPP and formulate testable hypotheses. We will begin with a set of basic hypotheses, which are not specific to the DPP but are commonplace in action theories in behavioral sciences:

³⁷The use of the general disposition for intuitive decision making is of limited relevance in the context of this experiment because the pressure and leisure conditions are explicitly designed to override this disposition.

Framing Hypothesis In the cooperative condition, participants contribute more to the PGG than in the neutral condition.

Prosociality Hypothesis The higher a participant's PSA score, the more she contributes to the PGG.

The first hypothesis is based on the Thomas Theorem (Thomas and Thomas 1928), which states that the interpretation of a situation determines the actions taken in response to that situation. However, a similar effect could also be derived from our DPP model, since in the cooperative condition the positive aspects of cooperation should be more focal and thus influence behavior in this direction.

The second hypothesis is rooted in the assumption that people have other-regarding preferences and behave accordingly (e.g., Andreoni 1995a; Fehr and Schmidt 1999). Again, this hypothesis can also be derived from the DPP model. In addition to the fact that the DPP allows for the rational evaluation of preferences through Type 2 processes, similar to the explanation just given, there is a second, DPP-specific mechanism: According to the DPP, more pronounced prosocial attitudes lead to an increased salience of those aspects that are consistent with these attitudes, such as the potential gains for other group members.

The DPP can also help to qualify both hypotheses: The principle of catalyzation suggests that the influence of priming and framing on behavior is expected to be stronger when the behavior is performed automatically and without conscious deliberation. Similarly, prosocial attitudes are expected to exert the greatest influence on behavior when they are processed automatically. Based on these theoretical considerations, the following two interaction hypotheses can be formulated:

Framing Interaction Hypothesis The effect of the priming and framing condition on contributions is greater in the pressure condition than in the leisure condition.

Prosociality Interaction Hypothesis The effect of the PSA score on contributions is greater in the pressure condition than in the leisure condition.

After formulating the experimental design and deriving the corresponding hypotheses, we turn to the empirical results of this study. Between 2017 and 2018, a total of 782 participants successfully participated in 64 sessions at the Leipziger Experimentallabor (LEx). Most participants were university students (87%) or former university students (6%), the average age was 26 years, and women were slightly overrepresented (66%). Participants earned an average of EUR 15.60 for the one-hour sessions.

Prior to examining the effects of time constraints on decision making, it is important to determine the extent to which participants adhered to our experimental protocol. Our results indicate a significant difference in mean response times between the two experimental conditions (pressure: $M = 8.3$ seconds, leisure: $M = 30.1$ seconds, $p < .001$, t-test), suggesting that our experimental intervention led to a greater occurrence of Type 1 reasoning among subjects in the pressure condition compared to those in the leisure condition.

Concerning the contributions to the PGG, nearly half of the participants chose to contribute either nothing, exactly half, or all of their endowment. On average, participants contributed 59.3% of their endowment. Regarding the PSA score, we observe a rather prosocial pool (see Figure 2(a) of contribution 5), with more than 93% of the participants having a PSA score higher than .5 resulting in an average PSA score of .676. Since we expected the strength of the effects of a Type 1 decision to differ depending on the PSA score, we assigned each participant to one of two categories: Participants have a high PSA score if their score is equal to or above the median ($= .659$), and they have an intermediate PSA score if it is below the median.

With regards to our two basic hypotheses on priming and framing as well as on prosocial attitudes, we find statistical support for both (see Figure 2.2).³⁸ Specifically, we observe that the contribution rate is significantly higher in the cooperative condition than in the neutral condition ($\Delta = .053$, $p = .014$). Furthermore, participants with a high PSA contribute on average 20 more than participants with

³⁸Unless otherwise specified, all further tests are Wilcoxon rank-sum tests. The stars in the graphs indicate the significance level of these tests: * $p < .05$, ** $p < .01$, *** $p < .001$.

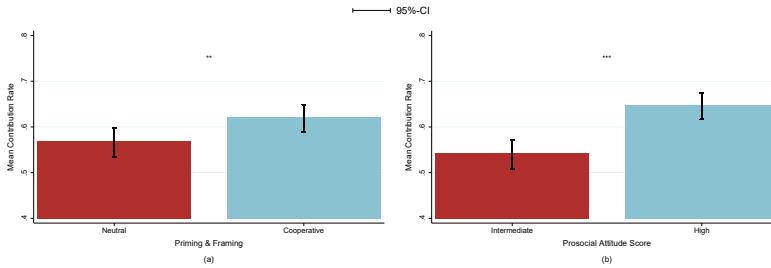


Figure 2.2: Bivariate effects on contribution in the PGG of (a) the priming and framing condition and (b) prosocial attitude score.

an intermediate PSA score ($\Delta = .105$, $p < .001$). The latter finding also holds when we treat the PSA score as quasi-metric and calculate the linear regression coefficient ($r = .515$, $p < .001$, OLS).

Let us now turn to the interaction effects of the pressure condition with either the PSA score or the priming and framing condition. In the context of the PSA score, we observe three results in Figure 2.3: First, in accordance with the prosociality hypothesis a high PSA score generally leads to a higher contribution; second, pressure does not significantly alter the decisions of participants with an intermediate PSA score ($\Delta = .021$, $p = .455$); and third, pressure significantly increases the contribution in the PSA high group ($\Delta = .069$, $p < .024$). This effect is even more pronounced when we exclude those participants who violated the time constraints of their respective time condition (remaining $N = 694$, PSA intermediate $\Delta = .001$, $p < .0910$, PSA high: $\Delta = .079$, $p < .012$). Overall, we find support for the prosociality interaction hypothesis.

With respect to the framing interaction hypotheses, we expect a similar pattern when substituting the PSA score with the priming and framing condition. However, in neither the neutral ($\Delta = .035$, $p < .336$) nor the cooperative ($\Delta = .005$, $p < .816$) condition does the time constraint have a significant effect on contribution, thereby refuting the prosociality interaction hypothesis.

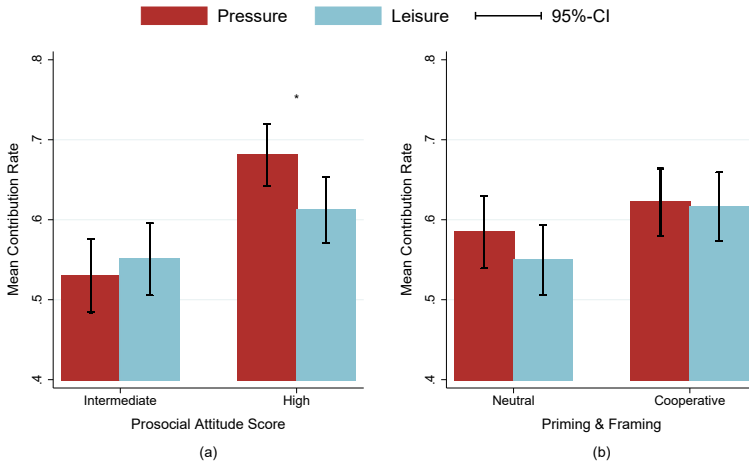


Figure 2.3: Interaction effects on contribution in the PGG between time constraints condition and (a) the prosocial attitude score as well as (b) the priming and framing condition.

Finally, we corroborate our previous findings through three multivariate linear regression models (see Table 2.3), with the participants' contribution in the PGG as the dependent variable. The first model tests our two basic hypotheses. The results indicate that both the cooperative condition ($p = .024$), as well as a high PSA score ($p < .001$), have a statistically significant positive impact on the contribution rate in the multivariate model.

The second model tests the more intricate prosociality interaction hypothesis. Compared to the first model, it also includes the pressure condition (as dummy variable) along with the interaction between a high PSA and the pressure condition. Just as in the previous model, we observe that the cooperative condition ($p = .028$) and a high PSA score ($p = .050$) have a positive and significant effect on contributions. Furthermore, and in accordance with the prosociality interaction hypothesis, we observe that pressure significantly affects the contribution of participants in the PSA high group ($p = .040$).

<i>Contribution</i>	Model 1	Model 2	Model 3
	Coefficient	Coefficient	Coefficient
Cooperative condition	.049* (.022)	.048* (.022)	.063* (.031)
High PSA score	.103*** (.022)	.060* (.030)	.104*** (.022)
Pressure condition		-.023 (.031)	.038 (.031)
Pressure × high PSA score		.089* (.043)	
Pressure × coop. condition			-.031 (.044)
Constant	.516*** (.019)	.528*** (.024)	.498*** (.024)
R^2	.035	.042	.037

Notes: $N = 782$, all variables are dummy variables with no = 0 and yes = 1.

OLS coefficients, standard errors in parentheses, * $p < .05$, ** $p < .01$, *** $p < .001$.

Table 2.3: Multivariate linear regression models of contribution in the PGG.

In the third model, we test the framing interaction hypothesis. This model is similar to the second, except that it uses a different interaction, namely that between the pressure condition and the framing condition. The general effects of the cooperative condition and a high PSA score are mostly comparable to those in the first model ($p = .039$ and $p < .001$, respectively). Regarding the interaction effects, we do not observe a significant change in contributions when participants in the cooperative condition are put under time pressure ($p = .483$).³⁹

Let us briefly assess our findings, beginning with the impact of the priming and framing condition. While we observed a statistically significant difference between the neutral and the cooperative conditions with respect to prosocial behavior, the effect size was smaller not only compared to the difference between participants

³⁹As a robustness check, the following control variables were included in all three models: Sex (m/f), age, age squared, previous lab experience (no/yes), incentive condition (k low/ k high), and time of PSA measurement (before/after the experiment). The effects not only remained robust in the presence of these control variables, but became even more pronounced. In addition, Model 2 and Model 3 were combined to test both interaction hypotheses simultaneously. In this case, the results are very similar to those of the individual models.

with intermediate and high levels of prosociality, but also smaller than the effects of framing reported in previous literature (e.g., Liberman et al. 2004). One potential explanation for the relatively modest effect observed in our study could be attributed to the possibility that the typical participants in such studies, predominantly students, already possess a cooperative mindset (framing) upon entering the laboratory experiment (cf. Bear and Rand 2016; Rand et al. 2012). As a consequence, the neutral condition would have led participants to maintain the cooperative framing they had brought with them, as they were not actively framed by us. Therefore, it would be interesting to investigate whether a more competitive framing would produce stronger direct as well as interaction effects.

In relation to the interaction between the pressure and cooperation conditions, our study found no significant effect, thus providing no empirical support for the framing interaction hypothesis. Nevertheless, this does not necessarily contradict the catalyzation principle, as the absence of a significant interaction effect may be attributed to the comparatively weak priming and framing effect. Indeed, various studies have corroborated that human actors tend to be more influenced by framing when they are required to make decisions more intuitively, either due to time constraints (Guo et al. 2017) or cognitive overload (Whitney et al. 2008). However, given the speculative nature of this explanation, further research is needed.

With respect to the effect of prosocial attitudes, our study provides empirical support for the principle of catalyzation in relation to cultural orientation stored in the nondeclarative memory. More precisely, our research suggests that time pressure serves as a moderating variable in the relationship between individuals' prosocial attitudes and their actual prosocial behavior. Specifically, our analysis revealed that participants with highly prosocial attitudes were more likely to engage in prosocial behavior. In addition, we observed that individuals with highly prosocial attitudes were more likely to engage in prosocial behavior under time pressure, whereas individuals with intermediate prosocial attitudes did not show a similar increase in prosocial behavior.

2.2.3 Status Characteristics Theory

The second approach under consideration is *Status Characteristics Theory* (SCT), which was originally proposed by Joseph Berger Berger and colleagues in the early 1970s to explain the emergence of hierarchies in small groups (Berger et al. 1972; Berger and Fisek 1970). Since then, SCT has undergone a steady process of empirical testing and theoretical development (e.g., Berger et al. 1977, 1985; Correll and Ridgeway 2003; Ridgeway and Kricheli-Katz 2013; Simpson et al. 2012). Although SCT stems from a different line of theoretical inquiry than the Dual Process Perspective (DPP), the ideas of SCT are in many ways compatible with the DPP (cf. Miles et al. 2019). In the following, we first provide a brief synopsis of SCT before highlighting its connections to the DPP.⁴⁰

SCT focuses on explaining the differential performance of group members at group tasks. For the theory to be applicable, the group task must meet two scope conditions: First, the group must be oriented toward accomplishing a specific group task; and second, the task must foster a collective orientation in which there is a general expectation that all group members are both able and willing to contribute to the group task (Webster and Driskell 1978, 222).

Given these scope conditions, SCT posits that observable status characteristics of actors influence behavior in group tasks by affecting performance expectations. The theory distinguishes between two types of status characteristics: diffuse and specific. While diffuse status characteristics, such as gender or ethnicity, affect performance expectations across a wide range of tasks, the influence of specific status characteristics, such as education, literacy, or knowledge in a particular domain, is restricted to a limited and well-defined range of settings (Correll and Ridgeway 2003). In a group task, these observable status characteristics become effective (or, in the language of SCT, *salient*) unless they are explicitly *dissociated* from the task (the so-called burden of proof assumption, Berger et al. 1972, 246).

In a group task, salient status characteristics influence expectations about the behavior of each group member, which are referred to as performance expectations

⁴⁰As more detailed discussion of SCT can be found in contribution 4 (Tutić and Grehl 2018).

(Berger et al. 1972). While a given status characteristic may have multiple manifestations, SCT typically works only with a binary evaluation of differences (e.g., competent vs. incompetent or educated vs. uneducated) in terms of performance expectations. However, whether a particular manifestation of a status characteristic is evaluated positively or negatively depends on the other manifestations of that status characteristic used for comparison. For example, a secondary education may evoke a different expectation when compared to a primary education than when compared to a tertiary education (Correll and Ridgeway 2003, 33).⁴¹

According to SCT, these expectations influence behavior via the following mechanisms: Actors compare their relative performance expectations with all others in the group. Actors who hold higher expectations of themselves relative to others are more likely to occupy a higher position in the power and prestige order of the group. Such an advantage in the power and prestige order, in turn, leads to greater opportunities to act as well as to an increased likelihood of accepting such opportunities. For example, if a particular status characteristic is associated with expectations of competence and usefulness of contributions to the group task, then individuals bearing this characteristic will occupy a higher position in the power and prestige order and, consequently, will be expected to contribute more to the group task.⁴²

Let us now briefly highlight the connections between SCT and the DPP, highlighting why SCT can be viewed as a narrow application of the general ideas of the

⁴¹Given that actors are typically judged by multiple status characteristics, such as gender, age, or education, these characteristics can elicit congruent or incongruent expectations. SCT contains a number of assumptions that deal with the combined effects of such combinations (e.g., Berger and Fisek 1970; Knottnerus and Greenstein 1981; Zelditch et al. 1980).

⁴²Note that several authors (Berger et al. 1972; Correll and Ridgeway 2003; Ridgeway and Kricheli-Katz 2013) emphasize that performance expectations are not always based solely on the fact that status characteristics are associated with actual competence in the group task. Instead, certain cascading effects may occur. For instance, suppose we observe actors in a group task, each actor having either the status characteristic X or Y. Now, if one of the two groups shows on average a better performance in the group task, even if only by mere chance, this observation may lead to the formation of status expectations for these characteristics. Over time, i.e., through multiple interactions, this effect may be further reinforced to the extent that group members are guided by expectations about these status characteristics. Thus, performance expectations function like self-fulfilling prophecies, with group members behaving in ways that confirm the expectations they hold for themselves and others (cf. Correll and Ridgeway 2003, 31).

DPP (cf. Miles et al. 2019). SCT concerns itself with cognition and its influences on actors' expectations and behavior. In doing so, it implicitly draws on the idea of automatic (intuitive) Type 1 reasoning, as actors recognize cues about the status of others and involuntarily derive expectations from them without conscious thought (Ridgeway and Kricheli-Katz 2013). At the same time, however, these intuitions can be overridden by the process of dissociation, where actors question whether a particular characteristic is relevant in the current situation. Thus, dissociation can be viewed as an intervening Type 2 process.

Despite this, SCT does not systematically address the distinction between intuition and deliberation. While there is some evidence in the literature suggesting that the processing of diffuse status characteristics operates exclusively via Type 1 reasoning, as these characteristics are universally and uncontestedly transformed into general expectations (Berger et al. 1977, 108f), the categorization of specific status characteristics is less clear. In contrast to diffuse status characteristics, specific status characteristics “carry cultural expectations for competence at limited, well-defined range of tasks and, consequently, only impact the formation of performance expectations in this limited range of settings” (Ridgeway and Kricheli-Katz 2013, 32). This assessment of relevance suggests that Type 2 processes are involved, at least in part, in the processing of specific state features, and thus that specific status characteristics require more deliberation than diffuse status characteristics.⁴³

Let us now turn to an empirical test of the validity of SCT. For this purpose, an experimental study is presented in the following section.

⁴³To the best of our knowledge, no one has done a systematic comparison or integration of the ideas presented in DPP and SCT. Thus, other authors may come to a different conclusion.

2.2.4 Status Characteristics and the Provision of Public Goods: Experimental Evidence

Against the background of this theoretical framework, this section presents an experimental study (outlined in contribution 4; Tutić and Grehl 2018) which tests the validity of the SCT with respect to the influence of social status on prosocial behavior in a group task.

To this end, we conducted a laboratory experiment in which participants interacted in a modified version of the Volunteer’s Timing Dilemma (VTD, Otsubo and Rapoport 2008; Weesie 1993). The VTD is a social dilemma in which a certain number of actors (i.e., the threshold) must volunteer and incur individual costs in order to produce a common good. The term *timing* refers to the experimental setting that the faster a group manages to do reach this threshold, the greater the common good produced. It is a dilemma because it creates tension among participants, for while there is an inherent incentive for each participant to contribute to the creation of the common good, there is also a simultaneous incentive to hold back if it appears that enough others are volunteering. Therefore, in situations where collusion is infeasible, the participants in a VTD find themselves without a dominant strategy (Diekmann 1985), which means that there is no single best course of action that an actor can take.

The VTD was used to examine the influence of high- and low-status characteristics on volunteering behavior. Prior to the VTD, participants were assigned to either the high-status group (designated as “stars”) or the low-status group (designated as “nonstars”). To emphasize this assignment, a small ceremony was held in which the stars received preferential treatment in the form of better seating arrangements, as well as beverages and chocolate. Status assignment was based on three experimental treatments: In the *random treatment*, which serves as the control group, participants were assigned to their respective groups purely by chance. In the *diffuse treatment*, participants were assigned on the basis of their self-reported subjective social status (Adler et al. 2000). Finally, in the *specific treatment*, the assignment was again purely random, but participants were told

that their assignment was based on their performance on a previously administered quiz.⁴⁴

The VTD was performed in groups of four, consisting of two stars and two nonstars, with at least two participants required to bear a private cost of $C = 40$ to produce a public good of $B = 100$. The potential size of the public good decreased by 1 unit per second, and if the group did not reach the threshold within 60 seconds, the public good was not produced. Participants played 15 rounds, and all matching was random and anonymous.

According to traditional game theory, the stable symmetric mixed Nash-equilibrium predicts that each participant has a probability of about 46% of choosing to volunteer and a probability of 54% of choosing not to volunteer.⁴⁵ However, traditional game theory does not provide any insight into how ascribed status may affect these probabilities. In contrast, SCT makes several predictions in this regard. The scope conditions of SCT are apparently met in the VTD, making it applicable to generate predictions. Simpson et al. (2012) derive several propositions from SCT for a collective action problem like the VTD. In contribution 4 we use these propositions to formulate specific hypotheses for this laboratory experiment, the two most important being:

Individual Initiative Hypothesis: Given that the status characteristic is not dissociated, high-status actors will contribute faster to the collective action than low-status actors.

Individual Contribution Hypothesis: Given that the status characteristic is not dissociated, high-status actors will contribute more resources to the collective action than low-status actors.

⁴⁴This quiz was said to be very relevant to the study to emphasize the importance of the skills demonstrated in it.

⁴⁵There is also a second symmetric mixed Nash-equilibrium with a positive probability of roughly 23%. However, this equilibrium is neither payoff-dominant nor stable (cf. Offerman et al. 2001), therefore, it is less likely that participants will coordinate on this equilibrium. Both predictions are obtained by solving the following equation: $u(\text{volunteer}) = B(1 - (1 - p)^3) - C = B(p^3 + 3 * (1 - p) * p^2) = u(\text{not volunteer})$. Finally, there is a third stable but also not payoff-dominant symmetric mixed Nash-equilibrium with $p = 0$.

In this study, status characteristics should be dissociated in the random treatment because the assignment was based on chance alone. On the contrary, according to SCT, status characteristics are not dissociated in either the diffuse or the specific treatment. Consequently, we hypothesized that in the diffuse and specific treatments, stars would both be more likely to contribute and do so more quickly than nonstars. In contrast, for the random treatment, we anticipated that there would be no discernible differences in the likelihood or timing of contributions between the groups.

The results of our laboratory experiment align with these predictions. In the control group, we observed no effect on either the likelihood or the speed with which stars volunteered for the VTD. However, in the experimental groups that received either diffuse or specific treatments, stars were both more likely and faster to volunteer than nonstars. These findings are robust with respect to multivariate models that include various control variables (Tutić and Grehl 2018, Table 1). In addition, we tested several other hypotheses based on the specific version of SCT proposed by Simpson et al. (2012) regarding the collective effect of status differentials. Specifically, Simpson et al. (2012) argue that groups with salient status characteristics that are not dissociated should be more effective and efficient. We thus hypothesized that groups in the experimental treatments should produce the collective good more often and with less loss than in the control treatment. Both hypotheses received at least weak statistical support.

In this contribution, we demonstrated that SCT can offer predictions that classical Rational Choice Theory cannot account for and that these predictions are supported by empirical data. Perhaps the most important aspect to consider is the comparison between the random and specific treatments. In both cases, the assignment of star status was randomized, but in the specific treatment, participants were made to believe that this assignment was based on specific knowledge. As a result, the variations in individual and group performance that we observed in this study cannot be attributed to differences in group composition. Rather, this finding illustrates the potential impact of status characteristics and the corresponding expectations they create.

The fact that these expectations are not always consciously formed or even perceived by the actors themselves can also be seen in the following finding: At the end of the experiment, participants completed a follow-up questionnaire. Among other things, they were asked whether the status assigned to them had significantly influenced their decisions in the VTD task (not reported in contribution 4). Surprisingly, even in the experimental treatments, not even 10% of the respondents agreed "somewhat" or "strongly" with this statement. Since this percentage is insufficient to account for the observed differences between treatments, it may indicate that the effects of status characteristics operated primarily at a subconscious and automatic level for a considerable portion of respondents.

2.2.5 Discussion

This section examined two alternative theoretical approaches, namely the Dual Process Perspective (DPP) and Status Characteristics Theory (SCT). For each approach, a laboratory experiment was conducted to ascertain its empirical validity. The research showed that both approaches can be used to derive sociologically relevant hypotheses about prosocial behavior. Our results, thus, demonstrate the theoretical and practical value of these approaches, highlighting their potential for further application and extension in future research.

A major disadvantage of SCT, however, is that it is limited in its application along two dimensions. First, the theory can only be meaningfully applied to situations for which the scope conditions are met, i.e., collective group tasks. Second, SCT focuses on a particular aspect of social relations, namely status characteristics. As a result, its applicability tends to be limited to situations in which status characteristics are not relevant, for example, because actors have no way of evaluating these characteristics. In contrast, DPP is capable of explaining a wider range of phenomena across various domains.

In addition, as already mentioned, SCT can be seen as a particular application of the DPP (cf. Miles et al. 2019). To appreciate this, one must consider that SCT proposes that individuals use automatic and unconscious social cognition to form

expectations about the competence and potential contributions of others in group tasks, which are based on observable status characteristics such as gender, race, and education. However, SCT recognizes that reliance on automatic processes is not absolute, as deliberative cognitive processes can intervene in expectation formation by explicitly dissociating a salient status characteristic from the group task. SCT thus illustrates the interaction between intuitive Type 1 processes, which are based on salient cues about status characteristics, and deliberative Type 2 processes, which can override some of the implications drawn by these intuitive processes.⁴⁶

2.3 External Validity and Further Applications

Given the limitations of the Status Characteristics Theory and the fact that it can also be understood as a specialized version of the Dual Process Perspective (DPP), we decided to focus our further research exclusively on the DPP. Building on the results of the laboratory experiments, we sought to determine the extent to which these results hold outside of controlled laboratory settings. The final two contributions of this thesis therefore address two important aspects that are relevant to sociological research. First, the question of external validity is examined, i.e., whether the results of the laboratory experiment can be generalized to real-world situations. Second, the integration of the concepts and insights of the DPP into traditional sociological research methods, such as surveys, is explored. To address these questions, a field experiment and an online survey are presented in the subsequent sections.

⁴⁶Although it is not entirely clear from the literature on SCT which processes are intuitive and which are deliberative. Nevertheless, this provides a promising avenue for further research, such as examining whether and how SCT can be integrated into DPP.

2.3.1 Dual Process and Prosocial Behavior in the Field

Based on the findings of the laboratory experiment discussed in section 2.2.2, contribution 5 (Grehl and Tutić 2022) investigates the external validity of these results. For this purpose, a non-reactive field experiment was conducted in which participants were unaware that they were part of a study. The use of such an experimental design has several advantages, including the elimination of the artificiality of a laboratory setting and the reduction of potential behavioral alterations of participants due to observation awareness (Baldassarri and Abascal 2017).

The field experiment used a variation of the lost letter technique (cf. Farrington and Knight 1979), in which participants receive a message that is obviously not intended for them. Specifically, six months after the initial laboratory experiment, the former participants were contacted via email through the official account of the experimental laboratory. The email was designed in a way that created a cover story to prevent the participants from being aware of the true purpose of the email. In order to make it obvious to the participants that the email was actually intended for another recipient, the first line of the content of the email was addressed to a person whose first and last name did not appear among the participants. The email thanked them for their participation in a study that had supposedly taken place the day before, even though they had not participated in a study for six months. Each participant was sent an unique payout code, which they could use to anonymously receive the money they had supposedly earned.⁴⁷ As a result, the participants had three possible courses of action: First, to report the alleged error (act prosocially); second, to claim the money for themselves (act proselfishly); or third, to simply do nothing (act neutrally). After two weeks, we ended the observation phase, disclosed our study to all participants, and invited them to participate in a follow-up survey.

By combining data from the field experiment, the follow-up survey, and the previous laboratory experiment, we aimed to examine whether the prosocial attitude

⁴⁷We also experimentally varied several details of the email text (see Grehl and Tutić 2022), but these are not relevant to the framing introduction.

previously measured in the laboratory experiment (the PSA score, see section 2.2.2) is relevant in a natural setting and whether the strength of this effect varies depending on the intuitiveness of the decision-making procedure. While the intuitiveness of the participants was manipulated experimentally in the lab, this was not feasible in the field. Instead, two measures of intuitiveness were employed. First, general intuitiveness, which reflects an individual's basic tendency to rely on intuition when making decisions, was assessed in the previous laboratory experiment using an extended version of the Cognitive Reflection Test (cf. Table 2.1). Second, the intuitiveness of the specific decision was measured in the follow-up survey through self-assessment (situational intuitiveness). Noteworthy, the two measures of intuitiveness were not significantly correlated.

These two measures were used to test whether they would act as a moderating factor in the association between prosocial attitudes and prosocial behavior. Specifically, similar to the laboratory experiment in section 2.2.2, we expected that the higher participants' prosocial attitudes scores, the more likely they would be to engage in prosocial behavior. Furthermore, and in accordance with the principle of catalyzation, we anticipated a positive interaction effect between prosocial attitudes and intuitiveness on prosocial behavior.

Our results suggest that the prosocial attitudes observed in the laboratory experiment can be generalized to a natural setting. That is, we found that participants who exhibited stronger prosocial attitudes were more likely to engage in helpful behavior and less likely to act selfishly. Moreover, and more importantly, we observed that the impact of prosocial attitudes on behavior was stronger among participants who relied more on intuition in their decision making, regardless of whether general or situational intuitiveness was taken into account.

In addition, we explored the question of whether increased intuitiveness in decision making alone leads to more prosocial behavior. This idea, which has come to be known as the intuitive prosociality hypothesis, has been raised in the context of the DPP in several studies (e.g., Rand et al. 2012; Zaki and Mitchell 2013). In this contribution, we argue that this hypothesis does not follow directly from the DPP, as an increase in intuition would only suggest a greater reliance on Type

1 processes, without implying any specific behavior. However, we note that the proposed effect of the intuitive prosociality hypothesis may still be observable in practice. This can happen when the study population is characterized by high levels of prosocial attitudes. In such a situation, the highly prosocial actors would be the driving force behind this effect, since, according to the catalyzation principle, they should behave more prosocially when they decide intuitively. Conversely, this means that we would not expect a pure effect of intuitiveness when we control for attitudes. This aligns precisely with the evidence gleaned from this study.

Overall, the results of this study are consistent with the theoretical tenets of the DPP and provide further empirical support for the principle of catalyzation. Furthermore, the demonstrated applicability of the DPP to real-world situations testifies to the practical value of this theoretical model. In sociological research, however, laboratory and field experiments are not the most common methods of analysis. As such, our subsequent focus will explore the feasibility of integrating the concepts of the DPP and principles within a frequently used and highly regarded technique of sociological data acquisition, namely surveys.

2.3.2 Dual Process and Voting Intentions

Finally, contribution 6 (Tutić and Grehl 2021) presents a novel application of the Dual Process Perspective (DPP) in the context of voting intentions. The study focuses on the influence of explicitly and implicitly measured attitudes on intentions to vote for a right-wing party. Specifically, the study investigates how these attitudes are interrelated and how intuitiveness, in conjunction with these attitudes, influences voting intentions. This exemplifies how the DPP can be applied to sociological research and highlights the benefits of distinguishing between the different cognitive processes that underlie behavior.

While we have already used the concept of attitudes in sections 2.2.2 and 2.3.1, in this contribution we further differentiate in this contribution between implicit and explicit attitudes. Implicit attitudes are evaluations that are deeply ingrained (to the point that the holder of these attitudes may not even be aware of them) but are

automatically and involuntarily activated by the attitude object (e.g., Fazio and Olson 2003; Wilson et al. 2000). This characterization suggests a close link between implicit attitudes and Type 1 processes. Indeed, the DPP postulates that implicit attitudes are cultural orientations that are stored in the nondeclarative memory system (Smith and DeCoster 2000, see also section 2.2.1). As such, they can be accessed by pure Type 1 processes in a relatively unbiased manner. Measuring implicit attitudes requires techniques that rely predominantly on Type 1 processes, such as the Implicit Association Test (IAT, Greenwald et al. 1998, 2003).

Explicit attitudes, on the other hand, are conscious, deliberate evaluations that are obtained through direct questioning of the respondents. For example, a common method is to ask respondents to indicate the extent to which they agree or disagree with a series of statements about an attitude object (e.g., Liebe and Beyer 2021). According to DPP, responses to such questions are based on implicit attitudes but may be biased by Type 2 processes based on motivational and situational considerations. For instance, individuals who implicitly reject refugees may change their explicit response based on personal convictions (“Rejecting refugees is bad”) or social desirability (“Others expect me not to reject refugees”). In this respect, explicit attitudes can be seen as distorted versions of “true” implicit attitudes. As a result, implicit and explicit attitudes may diverge significantly under certain conditions, and therefore, explicit attitudes may not be reliable proxies for implicit attitudes.⁴⁸

Contribution 6 investigates the association between implicit and explicit attitudes and their influence on voting intentions for the Alternative für Deutschland (AfD), the largest right-wing party in Germany. Previous studies have shown that opposition to refugees and support for populist ideas significantly increase the likelihood of supporting the AfD (cf. Hambauer and Mays 2018; Lengfeld 2017; Rippl and Seipel 2018). However, these studies only consider the explicit attitudes of the respondents. To address this shortcoming, we conducted a population-representative online survey to collect both implicit and explicit attitudes toward refugees (*racism*) and populist ideas (*populism*). In addition, the tendency of

⁴⁸In this respect, the DPP also draws on the literature on socially desirable responses (e.g., Paulhus 1984).

actors to rely on Type 1 reasoning was measured in order to explore potential interaction effects as proposed by the DPP and the principle of catalyzation (section 2.2.1).

Based on the preliminary discussion, we proposed the following hypotheses: First, the degree of respondents' tendency toward Type 1 intuitive reasoning is positively correlated with the strength of the association between their implicit and explicit attitudes. Second, explicit attitudes toward populism and racism have a positive impact on one's affinity toward the AfD. Third, and in accordance with the principle of catalyzation, individuals with a greater propensity for intuitive Type 1 reasoning are expected to demonstrate a stronger association between their implicit attitudes toward populism and racism and their affinity toward the AfD, as compared to their more reflective counterparts.

The empirical results show that implicit and explicit attitudes correlate only weakly (racism) or not at all (populism) when we do not control for the propensity toward reasoning type. This is in line with expectations, as Type 2 processes may introduce bias into explicit attitudes, especially if the questions used to measure these attitudes are considered sensitive. In terms of the functional relationship between implicit and explicit attitudes, a higher level of implicit attitudes toward racism or populism results in a significantly higher level of explicit attitudes toward the same issue. In addition, implicit attitudes have a stronger positive influence on explicit attitudes when actors tend to rely on Type 1 reasoning, although this finding is significant only for racist attitudes.

Regarding the voting intention for the AfD, which was measured by a classical voting intention question, we estimated a multivariate linear regression model that included all four types of attitudes as well as the tendency toward Type 1 reasoning and the interactions between this tendency and the two implicit attitudes.⁴⁹ Our results indicate that explicit attitudes toward racism and populism have a significant and positive impact on the inclination toward the AfD, which is in line with

⁴⁹In addition, the following control variables were included: age, gender, migration background, city/state, east/west Germany, education, class, employment status, occupational prestige, left-right self-assessment, and religiosity.

our expectations. In addition, we observe that implicit populism and a tendency toward Type 1 reasoning significantly increase the likelihood of declaring a voting intention for the AfD. Finally, we observe a positive interaction effect between Type 1 reasoning and both types of implicit attitudes, confirming the principle of catalyzation for both forms of attitudes. Notably, while explicit racism was the strongest attitudinal predictor of AfD voting intention in our model, the second most important attitudinal factor was implicit populism. Moreover, for intuitive respondents, a one standard deviation increase in implicit populism was even more important than other common indicators such as left-right self-placement or whether the respondent was from eastern Germany.⁵⁰

In summary, contribution 6 introduces an innovative application of the DPP, specifically focusing on the area of voting intentions. Specifically, the study examines the relationship between implicit and explicit attitudes and their impact on the voting intentions for the AfD. The results indicate that, beyond explicit attitudes, implicit attitudes toward racism and populism positively influence inclination toward the AfD, and that this effect is mediated by an individual's tendency to engage in intuitive Type 1 reasoning. These findings highlight the usefulness of employing DPP concepts in sociological research, and the potential benefits of further exploring this approach to better understand the complexities of attitudes and voting behavior. Moreover, the study demonstrates that DPP concepts can be integrated into standard surveys commonly used in sociology.

2.3.3 Discussion

In this section, we presented a non-reactive field experiment and an online survey to assess the potential of the Dual Process Perspective (DPP) for sociological research beyond the controlled setting of laboratory experiments. Our goal was to test whether meaningful hypotheses could be derived from the concepts and ideas of the DPP and whether they would be confirmed empirically.

⁵⁰The AfD is particularly strong in eastern Germany, where its share of the vote in elections is three to four times higher than in other parts of Germany.

Specifically, in the field experiment, we observed that both a general tendency to rely on intuitive reasoning, as measured by the Cognitive Reflection Test, and situational intuitiveness, as measured by self-report, acted as moderators of the relationship between prosocial attitudes and prosocial behavior, as proclaimed by the catalyzation principle. Interestingly, we found that the two measures were not significantly correlated, suggesting that they are not mere substitutes and that each plays a unique role in shaping behavior. However, the exact nature of the functional relationship between these measures remains uncertain and requires further empirical investigation to better understand the mechanisms underlying their impact on behavior.

In the online survey, we found that incorporating implicit attitudes may provide a potential advantage over the sole use of explicitly measured attitudes. Our results suggest that implicit attitudes may play an important role in shaping voting intentions, especially for individuals who rely more on intuition. Of particular note is the fact that implicit attitudes toward populism had a stronger effect on voting for the right-wing AfD party than explicit attitudes toward populism. And that for intuitive individuals, implicit populism was even more important than other common indicators of AfD inclination.

Overall, our results suggest that the DPP can make sociologically relevant contributions beyond the laboratory setting. By using the DPP, we were able to derive meaningful hypotheses about real-world behavior or intentions toward such behavior, and these hypotheses were empirically supported in both our field experiment and online survey. This underscores the value of the DPP in providing meaningful insights into human behavior in real-world contexts.

3 Conclusion

This thesis represents a scientific journey aimed at investigating the role of cognition in human action. Central to this investigation is an effort to unravel the cognitive mechanisms that drive decision making and the resulting actions. Grounded in dissatisfaction with traditional Rational Choice Theory (RCT), which is based on oversimplifying assumptions, especially with respect to the cognition of actors (cf. contribution 1), the objective of this thesis is to provide methodological guidance for the development of an alternative theory of action that more accurately reflects the actual decision-making processes of human decision makers.

As part of this ambitious endeavor, the first step was to gain a more granular understanding of actors' cognitive abilities, particularly with respect to iterated reasoning. Findings from contributions 2 and 3 reveal that human actors display significant constraints in their iterative reasoning abilities, and that they are well aware of these limitations in others and tend to adapt their behavior accordingly, provided that the problems at hand are not overly complex. At the same time, these contributions shed light on the fact that reasoning style, i.e., the tendency of actors to make decisions more intuitively or deliberatively, exerts a substantial influence on their performance. Remarkably, there was even evidence suggesting that the variance in reasoning style might be more important than the typical variance in cognitive ability. This finding triggered a pivotal shift in the focus of this research from cognitive ability to reasoning style.

As a result of this shift, this thesis also turned to theoretical approaches that either explicitly or implicitly address the distinction between intuitive and deliberative decision making. Specifically, the Dual Process Perspective (DPP) and Status Characteristics Theory (SCT) were examined. Drawing upon these theoretical frameworks, we derived hypotheses in the context of prosocial behavior and tested

them in two separate laboratory experiments. The main tenets of both approaches were by and large empirically confirmed in the laboratory.

In light of the promising results from the laboratory experiments, the final part of this thesis ventured to examine the DPP beyond the controlled environment of the laboratory. This research effort focused exclusively on the DPP due to its broader theoretical scope in contrast to the SCT. The subsequent empirical research demonstrated the effectiveness of the DPP in the real-life context of a field experiment, as well as its compatibility with a common research methodology in the social sciences, namely surveys. The successful application of the DPP under more realistic conditions and its potential for integration with traditional research methods further underscores its versatility as a powerful tool for understanding and predicting social action.

Even though the DPP has proven to be a useful tool, it is important to recognize and discuss some of its potential limitations, particularly with regard to the quality criteria of a theory of action discussed earlier in section 1.3. The first criticism that may be encountered relates to the binary representation of human cognition in the DPP, which may not be an accurate representation of reality (e.g., Leschziner and Brett 2019; van Bavel et al. 2012). Some scholars argue that there may be more than two types of processes (cf. Evans 2009), while others suggest that there may be only one (e.g., Erb et al. 2003). In response to this objection, it is vital to remember that while the question of the “true” structure of the human mind is of course an important one, it is one that should be answered by cognitive scientists rather than sociologists. As such, sociologists should be open to the insights of this field of research and seek to develop models that do justice to both cognitive reality and the genuinely sociological demands of a sociological theory of action. In essence, the two-process conception in the DPP should be seen as a heuristic tool: As long as it can be used to derive interesting predictions that are empirically confirmed, it can be utilized until a more accurate theory replaces it.

Another objection to the DPP is that it is less parsimonious than RCT because it relies on a greater number of concepts, such as the reasoning style of actors. Therefore, in order to exploit the full potential of the DPP, more detailed measures

are required. However, even without such information, the DPP is still applicable to most settings, providing a significant information content. For example, it is still possible to formulate behavioral predictions based on the assumption that everyone decides purely intuitively or purely deliberatively. If the theory is to be validated, then the observed behavior must lie between these two extremes. Moreover, as demonstrated in the laboratory experiment on DPP (section 2.2.2), one can experimentally manipulate the situational reasoning style of actors, thus reducing the need to measure it. Nevertheless, future research should focus on identifying proxies as indicators of implicit attitudes and the tendency to use either intuitive or deliberative reasoning styles. This could lead to a new line of research exploring whether certain social groups are more likely to engage in intuitive or deliberative reasoning, potentially shedding light on new forms of social inequality (cf. Brett and Miles 2021).⁵¹ In addition, the application of novel techniques such as computer-assisted natural language processing can facilitate the use of process-generated textual data, for instance, to measure implicit attitudes by analyzing newspaper articles or social media posts (Bhatia and Walasek 2023).

An additional point of criticism could be that the DPP lacks the same level of axiomatic foundation or formalization as the classical RCT. Although this may undermine its precision in comparison to RCT, this point is qualified by the fact that, apart from RCT and its derivatives, most sociological theories of action also lack such an axiomatic foundation. However, there have been efforts to develop an axiomatic characterization for some of the ideas of the DPP (e.g., Tutić 2015b), indicating a continuous pursuit of theoretical rigor and refinement within this area of study.

To outline the direction of future research, it should be noted that although this work has demonstrated the potential of the DPP for sociological research, there are several empirical and theoretical challenges associated with the framework that remain to be resolved. For example, while the mechanisms such as the conditions

⁵¹A similar discourse is already underway about the unequal distribution of self control in society (e.g., Kroneberg and Schulz 2018) and the implications for life chances (e.g., Daly et al. 2015; Nakhaie et al. 2000). Self control can thus be seen as a deliberative process that overrides impulsive defaults.

for the intervention of Type 2 processes have been roughly outlined, further refinement is needed. To this end, future research should focus on a careful examination of both the theoretical intricacies and the empirical dimensions of the DPP. This should include further investigation of the specific factors that trigger Type 2 processes, the interplay between Type 1 and Type 2 processes, or how social and situational factors affect these cognitive processes.

Another promising area for future research is the transition from an individual-focused theory of action to a full-fledged sociological theory capable of explaining complex macro-phenomena. Predicting societal outcomes at the macro level requires a theory of aggregation, as indicated by the micro-macro link (see section 1.2). Despite the valuable insights provided by the DPP at the individual level, its potential for application at the macro level remains largely unexplored. Using the model presented in this thesis, it is only viable to aggregate behavior under the assumption that each actor's behavior is independent of others. However, many macro-level social phenomena require more complex aggregation rules, such as in cases where actors are influenced by past behavior or by their expectations of the behavior of others (cf. Ylikoski 2021). Once the DPP model has been given a solid theoretical as well as empirical foundation, and the ongoing debates about the specific interplay between the different types of processes have been settled, its integration into such more complex aggregation rules would be the next logical step. With this in mind, we propose two possible trajectories for the further development of the DPP in this context.

The first potential avenue is closely related to classical game theory. By integrating the fundamental principles of DPP into game-theoretic models, new equilibrium concepts could be formulated. Two interesting approaches presented in contribution 1, namely the procedural rational equilibrium (Osborne and Rubinstein 1998) and the level-k model (Bosch-Domènech et al. 2002; Costa-Gomes and Crawford 2006; Diekmann 2009), could be seen as representatives of such an endeavor. For example, the level-k model acknowledges the diversity of different reasoning styles among actors, suggesting that some individuals may employ an intuitive level-0 strategy, while others may resort to a more deliberative strategy. Notably, this

approach goes beyond merely identifying equilibria and can also be used to predict the possible evolutionary paths to equilibrium states from non-equilibrium starting points (cf. Diekmann 2009).

A second conceivable route for development is to incorporate the principles of DPP into models of social dynamics. These models, as put forward by analytical sociology (cf. Hedström and Bearman 2009; Manzo 2021), seek to explain collective outcomes by explicitly modeling the interactions of the actors involved. Using techniques such as agent-based models (Manzo 2022), they can easily accommodate alternative formulations of the cognitive underpinnings of the actors involved and thus arrive at new and interesting predictions (e.g., Bear and Rand 2016). To consider a concrete application, one can think of Social Learning Theory. Researchers in this field have identified several stylized facts about social learning biases, such as the status bias or the confirmation bias, that influence the extent to which individuals adopt an opinion in a social setting (Mesoudi 2011). However, the consistent empirical recognition of these biases contrasts with an apparent lack of theoretical frameworks that provide clarity on their formation as well as function at a deeper level. Here, the DPP could not only provide a theoretical underpinning for these biases by rooting them in cognitive processes, but also pave the way for novel hypotheses regarding their manifestation and impact.

In summary, this thesis underscores the pivotal role that cognition can play in shaping decision making and, consequently, action. While this thesis does not claim that cognitive factors are universally indispensable to all sociological analyses, it does suggest that their inclusion can enhance theoretical insights and improve predictions. One of the key insights of this research is the importance of the distinction between intuitive and deliberative reasoning styles and the role of their interaction as crucial components of a more comprehensive theory of action. In this context, the DPP represents a particularly promising theoretical approach that provides tangible concepts of human cognition without resorting to intricate measurement techniques. At the same time, the DPP facilitates the formulation of novel explanatory mechanisms and the derivation of intriguing hypotheses. This balance of applicability and theoretical depth makes it a valuable tool for ad-

vancing sociological inquiry. Moreover, the DPP provides a basic framework for bridging different sociological perspectives, such as practice theory and RCT. By synthesizing these diverse ideas, the DPP can help develop a more nuanced theory of action that takes into account the full potential of sociological ideas (cf. Tutić 2022; Vila-Henninger 2021).

Finally, the study of the complex interplay between cognition and social action is an ongoing enterprise. It is our hope that this thesis, or parts of it, will make a meaningful contribution to that endeavor.

Bibliography

- Abbate, C. S., S. Boca, and S. Ruggieri (2013). The effect of prosocial priming in the presence of bystanders. *The Journal of Social Psychology* 153(5), 619–622.
- Abell, P., T. Felin, and N. Foss (2008). Building micro-foundations for the routines, capabilities, and performance links. *Managerial and Decision Economics* 29(6), 489–502.
- Abraham, M. and T. Voss (2000). Rational choice theory in sociology: A survey. In S. R. Quah and A. Sales (Eds.), *The International Handbook of Sociology*, pp. 50–83. London & Thousand Oaks: Sage Publications.
- Adcock, R. and D. Collier (2001). Measurement validity: A shared standard for qualitative and quantitative research. *American Political Science Review* 95(3), 529–546.
- Adler, N. E., E. S. Epel, G. Castellazzo, and J. R. Ickovics (2000). Relationship of subjective and objective social status with psychological and physiological functioning: Preliminary data in healthy, white women. *Health Psychology* 19(6), 586–592.
- Agranov, M., A. Caplin, and C. Tergiman (2015). Naive play and the process of choice in guessing games. *Journal of the Economic Science Association* 1(2), 146–157.
- Allais, M. (1979). The so-called allais paradox and rational decisions under uncertainty. In M. Allais and O. Hagen (Eds.), *Expected Utility Hypotheses and the*

- Allais Paradox: Contemporary Discussions of the Decisions under Uncertainty with Allais' Rejoinder*, pp. 437–681. Dordrecht: Springer.
- Alós-Ferrer, C. and M. Garagnani (2020). The cognitive foundations of cooperation. *Journal of Economic Behavior and Organization* 175, 71–85.
- Andreoni, J. (1995a). Cooperation in public-goods experiments: Kindness or confusion? *The American Economic Review* 85(4), 891–904.
- Andreoni, J. (1995b). Warm-glow versus cold-prickle: The effects of positive and negative framing on cooperation in experiments. *Quarterly Journal of Economics* 110(1), 1–21.
- Anscombe, G. E. M. (2000 [1957]). *Intention*. Cambridge: Harvard University Press.
- Arad, A. and A. Rubinstein (2012). The 11–20 money request game: A level-k reasoning study. *American Economic Review* 102(7), 3561–3573.
- Aumann, R. J. (1976). Agreeing to disagree. *Annals of Statistics* 4(6), 1236–1239.
- Aumann, R. J. (1985). What is game theory trying to accomplish? In K. J. Arrow and S. Honkapohja (Eds.), *Frontiers of Economics*, pp. 5–44. Oxford: Basil Blackwell.
- Aumann, R. J. (1995). Backward induction and common knowledge of rationality. *Games and Economic Behavior* 8(1), 6–19.
- Aumann, R. J. (1998). On the centipede game. *Games and Economic Behavior* 23(1), 97–105.
- Aumann, R. J. and A. Brandenburger (1995). Epistemic conditions for nash equilibrium. *Econometrica* 63(5), 1161–1180.
- Aumann, R. J. and S. Sorin (1989). Cooperation and bounded recall. *Games and Economic Behavior* 1(1), 5–39.
- Baddeley, A. D. (1968). A 3 min reasoning test based on grammatical transformation. *Psychonomic Science* 10(10), 341–342.

- Baldassarri, D. and M. Abascal (2017). Field experiments across the social sciences. *Annual Review of Sociology* 43, 41–73.
- Bandalos, D. L. (2018). *Measurement Theory and Applications for the Social Sciences*. London: Guilford Publications.
- Bargh, J. A. (2006). What have we been priming all these years? On the development, mechanisms, and ecology of nonconscious social behavior. *European Journal of Social Psychology* 36(2), 147–168.
- Batson, C. D. and A. A. Powell (2003). Altruism and prosocial behavior. In T. Millon and I. B. Weiner (Eds.), *Handbook of Psychology*, pp. 463–484. Hoboken: Wiley.
- Baudson, T. G. and F. Preckel (2016). Mini-q: Intelligenzscreening in drei Minuten. *Diagnostica* 62(3), 182–197.
- Bayer, R.-C. and M. Chan. The dirty faces game revisited. In A. Rambaldi (Ed.), *Proceedings of ESAM07*, pp. 1–26. Brisbane: University of Queensland.
- Bear, A. and D. G. Rand (2016). Intuition, deliberation, and the evolution of cooperation. *Proceedings of the National Academy of Sciences* 113(4), 936–941.
- Becker, G. S. (1968). Crime and punishment: An economic approach. *Journal of Political Economy* 76(2), 169–217.
- Becker, G. S. (1971). *Economic Theory*. New York: Knopf.
- Berger, J., B. P. Cohen, and M. Zelditch Jr (1972). Status characteristics and social interaction. *American Sociological Review* 37(3), 241–255.
- Berger, J. and M. H. Fisek (1970). Consistent and inconsistent status characteristics and the determination of power and prestige orders. *Sociometry* 33(3), 287–304.
- Berger, J., M. H. Fisek, R. Z. Norman, and M. Zelditch, Jr. (1977). *Status Characteristics and Social Interaction: An Expectation-States Approach*. New York: Elsevier.

- Berger, J., D. G. Wagner, and M. Zelditch, Jr. (1985). Introduction: Expectation states theory - review and assessment. In J. Berger and M. Zelditch, Jr. (Eds.), *Status, Rewards, and Influence: How Expectations Organize Behaviour*, pp. 1–72. San Francisco: Jossey-Bass.
- Bhatia, S. and L. Walasek (2023). Predicting implicit attitudes with natural language data. *Proceedings of the National Academy of Sciences* 120(25), e2220726120.
- Bierhoff, H. W. (2008). Prosocial behaviour. In M. Hewstone, W. Stroebe, and K. Jonas (Eds.), *An Introduction to Social Psychology*, pp. 176–195. Oxford: Blackwell.
- Binmore, K. (1999). Why experiment in economics? *The Economic Journal* 109(453), 16–24.
- Binmore, K. G. (1996). A note on backward induction. *Games and Economic Behavior* 17, 135–137.
- Binmore, K. G. (2007). *Does Game Theory Work? The Bargaining Challenge*. Cambridge: MIT Press.
- Binmore, K. G. and A. Brandenburger (1990). Common knowledge and game theory. In K. G. Binmore (Ed.), *Essays on the Foundation of Game Theory*, pp. 105–150. Oxford: Basil Blackwell.
- Blau, P. M. (1977). A macrosociological theory of social structure. *American Journal of Sociology* 83(1), 26–54.
- Blumer, H. (1986). *Symbolic Interactionism: Perspective and Method*. Berkeley and Los Angeles: University of California Press.
- Bosch-Domènech, A., J. G. Montalvo, R. Nagel, and A. Satorra (2002). One, two, (three), infinity, ... newspaper and lab beauty-contest experiments. *American Economic Review* 92(5), 1687–1701.
- Bourdieu, P. (1984). *Distinction: A Social Critique of the Judgement of Taste*. Cambridge: Harvard University Press.

- Bourdieu, P. (1990). *The Logic of Practice*. Stanford: Stanford University Press.
- Brañas-Garza, P., T. Garcia-Muñoz, and R. H. González (2012). Cognitive effort in the beauty contest game. *Journal of Economic Behavior and Organization* 83(2), 254–260.
- Brandenburger, A. and E. Dekel (1993). Hierarchies of beliefs and common knowledge. *Journal of Economic Theory* 59(1), 189–198.
- Braun, N. (2008). Theorie in der Soziologie. *Soziale Welt* 59(4), 373–395.
- Breen, R. and J. H. Goldthorpe (1997). Explaining educational differentials: Towards a formal rational action theory. *Rationality and Society* 9(3), 275–305.
- Brett, G. and A. Miles (2021). Who thinks how? Social patterns in reliance on automatic and deliberate cognition. *Sociological Science* 8, 96–118.
- Brocas, I. and J. D. Carrillo (2014). Dual-process theories of decision-making: A selective survey. *Journal of Economic Psychology* 41, 45–54.
- Bunge, M. (1996). *Finding Philosophy in Social Science*. New Haven: Yale University Press.
- Buskens, V. and W. Raub (2013). Rational choice research on social dilemmas: Embeddedness effects on trust. In R. Wittek, T. A. B. Snijders, and V. Nee (Eds.), *The Handbook of Rational Choice Social Research*, pp. 113–150. Stanford: Stanford University Press.
- Camerer, C. F. (2003). *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton: Princeton University Press.
- Capra, C. M., J. K. Goeree, R. Gomez, and C. A. Holt (1999). Anomalous behavior in a traveler’s dilemma? *American Economic Review* 89(3), 678–690.
- Carpenter, J., M. Graham, and J. Wolf (2013). Cognitive ability and strategic sophistication. *Games and Economic Behavior* 80, 115–130.
- Chaiken, S. and Y. Trope (Eds.) (1999). *Dual-Process Theories in Social Psychology*. New York: Guilford Press.

- Chou, E., M. McConnell, R. Nagel, and C. R. Plott (2009). The control of game form recognition in experiments: Understanding dominant strategy failures in a simple two person “guessing” game. *Experimental Economics* 12(2), 159–179.
- Coleman, J. S. (1986). Social theory, social research, and a theory of action. *American Journal of Sociology* 91(6), 1309–1335.
- Coleman, J. S. (1987). Microfoundations and macrosocial behavior. In J. C. Alexander, B. Giesen, R. Münch, and N. J. Smelser (Eds.), *The Micro-Macro Link*, pp. 153–173. Berkeley and Los Angeles: University of California Press.
- Coleman, J. S. (1990). *Foundations of Social Theory*. Cambridge, Mass: Belknap Press.
- Coleman, J. S. (1993). The rational reconstruction of society: 1992 presidential address. *American Sociological Review* 58(1), 1–15.
- Cornish, D. B. and R. V. Clarke (2014). *The Reasoning Criminal: Rational Choice Perspectives on Offending*. New Brunswick: Springer.
- Correll, S. J. and C. L. Ridgeway (2003). Expectation states theory. In John Delamater (Ed.), *Handbook of Social Psychology*, pp. 29–53. New York: Kluwer Academic/Plenum Publishers.
- Costa-Gomes, M. A. and V. P. Crawford (2006). Cognition and behavior in two-person guessing games: An experimental study. *American Economic Review* 96(5), 1737–1768.
- Crawford, V. P. and N. Iriberri (2007). Level-k auctions: Can a nonequilibrium model of strategic thinking explain the winner’s curse and overbidding in private-value auctions? *Econometrica* 75(6), 1721–1770.
- Daly, M., L. Delaney, M. Egan, and R. F. Baumeister (2015). Childhood self-control and unemployment throughout the life span: Evidence from two British cohort studies. *Psychological Science* 26(6), 709–723.

- Darley, J. M. and B. Latané (1968). Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology* 8(4), 377–383.
- De Neys, W. (2006). Dual processing in reasoning: Two systems but one reasoner. *Psychological Science* 17(5), 428–433.
- Devetag, G. and M. Warglien (2003). Games and phone numbers: Do short-term memory bounds affect strategic behavior? *Journal of Economic Psychology* 24(2), 189–202.
- Diekmann, A. (1985). Volunteer's dilemma. *Journal of Conflict Resolution* 29(4), 605–610.
- Diekmann, A. (1992). Soziale Dilemmata. Modelle, Typisierungen und empirische Resultate. In H.-J. Andreß, J. Huinink, H. Meinken, D. Rumianek, W. Sodeur, and G. Sturm (Eds.), *Theorie, Daten, Modelle*, pp. 177–203. München: Oldenbourg Verlag.
- Diekmann, A. (2004). The power of reciprocity: Fairness, reciprocity, and stakes in variants of the dictator game. *Journal of Conflict Resolution* 48(4), 487–505.
- Diekmann, A. (2009). Rational choice, evolution and the beauty contest. In M. Cherkaoui and P. Hamilton (Eds.), *Raymond Boudon - a Life in Sociology*, pp. 195–206. Oxford: The Bardwell Press.
- Diekmann, A. and T. Voss (2004). Die theorie rationalen handelns. stand und perspektiven. In A. Diekmann and T. Voss (Eds.), *Rational-Choice-Theorie in den Sozialwissenschaften: Anwendungen und Probleme*, pp. 13–29. München: Oldenbourg Verlag.
- DiMaggio, P. (1997). Culture and cognition. *Annual Review of Sociology* 23(1), 263–287.
- Dufwenberg, M., R. Sundaram, and D. J. Butler (2010). Epiphany in the game of 21. *Journal of Economic Behavior and Organization* 75(2), 132–143.

- Duggan, M. and S. D. Levitt (2002). Winning isn't everything: Corruption in sumo wrestling. *American Economic Review* 92(5), 1594–1605.
- Eagly, A. H. and S. Chaiken (1993). *The Psychology of Attitudes*. San Diego:: Harcourt Brace Jovanovich.
- Elster, J. (Ed.) (1999). *Addiction*. Entries and Exits. Russell Sage Foundation.
- Elster, J. (2007). *Explaining Social Behavior: More Nuts and Bolts for the Social Sciences*. Cambridge: Cambridge University Press.
- Englich, B., T. Mussweiler, and F. Strack (2006). Playing dice with criminal sentences: The influence of irrelevant anchors on experts' judicial decision making. *Personality and Social Psychology Bulletin* 32(2), 188–200.
- Erb, H.-P., A. Kruglanski, W. Y. Chun, A. Pierro, L. Mannetti, and S. Spiegel (2003). Searching for commonalities in human judgement: The parametric uni-model and its dual mode alternatives. *European Journal of Social Psychology* 14(1), 1–47.
- Esser, H. (1993). *Soziologie: Allgemeine Grundlagen*. Frankfurt a.M.: Campus.
- Esser, H. (1996). Die Definition der Situation. *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 48(1), 1–34.
- Esser, H. (2010). Das Modell der Frame-Selektion. Eine allgemeine Handlungstheorie für die Sozialwissenschaften? In G. Albert and S. Sigmund (Eds.), *Soziologische Theorie kontrovers*. Sonderheft 50/2010 der *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, pp. 45–62. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Esser, H. and C. Kroneberg (2015). An integrative theory of action: The model of frame selection. In E. J. Lawler, S. R. Thye, and J. Yoon (Eds.), *Order on the Edge of Chaos: Social Psychology and the Problem of Social Order*, pp. 63–85. Cambridge: Cambridge University Press.

- Evans, J. S. B. T. (2009). How many dual-process theories do we need? One, two, or many? In J. S. B. T. Evans and K. Frankish (Eds.), *In Two Minds*, pp. 33–54. Oxford: Oxford University Press.
- Evans, J. S. B. T. (2010). Intuition and reasoning: A dual-process perspective. *Psychological Inquiry* 21(4), 313–326.
- Evans, J. S. B. T. (2011). Dual-process theories of reasoning: Contemporary issues and developmental applications. *Developmental Review* 31(2-3), 86–102.
- Evans, J. S. B. T. and K. Frankish (Eds.) (2009). *In Two Minds: Dual Processes and Beyond*. Oxford: Oxford University Press.
- Evans, J. S. B. T. and K. E. Stanovich (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science* 8(3), 223–241.
- Farrington, D. P. and B. J. Knight (1979). Two non-reactive field experiments on stealing from a 'lost' letter. *British Journal of Social and Clinical Psychology* 18(3), 277–284.
- Fazio, R. H. (1990). Multiple processes by which attitudes guide behavior: The mode model as an integrative framework. *Advances in Experimental Social Psychology* 23, 75–109.
- Fazio, R. H. (1995). Attitudes as object-evaluation associations: Determinants, consequences, and correlates of attitude accessibility. In R. E. Petty and J. A. Krosnick (Eds.), *Attitude Strength: Antecedents and Consequences*, pp. 247–282. New Jersey: Lawrence Erlbaum Associates.
- Fazio, R. H. and M. A. Olson (2003). Implicit measures in social cognition research: Their meaning and use. *Annual Review of Psychology* 54, 297–327.
- Fehr, E. and K. M. Schmidt (1999). A theory of fairness, competition and cooperation. *Quarterly Journal of Economics* 114(3), 817–868.

- Fey, M., R. D. McKelvey, and T. R. Palfrey (1996). An experimental study of constant-sum centipede games. *International Journal of Game Theory* 25(3), 269–287.
- Frank, R. H. (1988). *Passions Within Reason: The Strategic Role of the Emotions*. New York: W. W. Norton.
- Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives* 19(4), 25–42.
- Friedman, D. and M. Hechter (1988). The contribution of rational choice theory to macrosociological research. *Sociological Theory* 6(2), 201–218.
- Friedman, M. (Ed.) (1953). *Essays in Positive Economics*. Chicago: University of Chicago Press.
- Furnham, A. and H. C. Boo (2011). A literature review of the anchoring effect. *The Journal of Socio-Economics* 40(1), 35–42.
- Geanakoplos, J. (1992). Common knowledge. *Journal of Economic Perspectives* 6(4), 53–82.
- Geanakoplos, J. (2021). Game theory without partitions, and applications to speculation and consensus. *The BE Journal of Theoretical Economics* 21(2), 361–394.
- Georganas, S., P. J. Healy, and R. A. Weber (2015). On the persistence of strategic sophistication. *Journal of Economic Theory* 159, 369–400.
- Giddens, A. (1984). *The Constitution of Society: Outline of the Theory of Structuration*. Berkeley: University of California Press.
- Gilbert, D. T. (1999). What the mind’s not. In S. Chaiken and Y. Trope (Eds.), *Dual-Process Theories in Social Psychology*, pp. 3–11. New York: Guilford Press.
- Gneezy, U., A. Rustichini, and A. Vostroknutov (2010). Experience and insight in the race game. *Journal of Economic Behavior and Organization* 75(2), 144–155.

- Goeree, J. K. and C. A. Holt (2001). Ten little treasures of game theory and ten intuitive contradictions. *American Economic Review* 91(5), 1402–1422.
- Goffman, E. (1974). *Frame Analysis: An Essay on the Organization of Experience*. New York: Harper & Row.
- Gossen, H. H. (1983 [1854]). *The Laws of Human Relations and the Rules of Human Action Derived Therefrom*. Cambridge, MA and London: The MIT Press.
- Granovetter, M. (1985). Economic action and social structure: The problem of embeddedness. *American Journal of Sociology* 91(3), 481–510.
- Grayot, J. D. (2020). Dual process theories in behavioral economics and neuroeconomics: A critical review. *Review of Philosophy and Psychology* 11(1), 105–136.
- Greene, J. D., S. A. Morelli, K. Lowenberg, L. E. Nystrom, and J. D. Cohen (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition* 107(3), 1144–1154.
- Greenwald, A. G., D. E. McGhee, and Schwartz, Jordan L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology* 74(6), 1464–1480.
- Greenwald, A. G., B. A. Nosek, and M. R. Banaji (2003). Understanding and using the implicit association test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology* 85(2), 197.
- Grehl, S. (2020). Verhaltensökonomik und begrenzte Rationalität. In A. Tutić (Ed.), *Rational Choice*, pp. 150–178. Oldenbourg: De Gruyter.
- Grehl, S. and A. Tutić (2015). Experimental evidence on iterated reasoning in games. *PLoS ONE* 10(8), e0136524.
- Grehl, S. and A. Tutić (2022). Intuition, reflection, and prosociality: Evidence from a field experiment. *PLoS ONE* 17(2), e0262476.

- Guo, L., J. S. Trueblood, and A. Diederich (2017). Thinking fast increases framing effects in risky decision making. *Psychological Science* 28(4), 530–543.
- Hambauer, V. and A. Mays (2018). Wer wählt die AfD? – Ein Vergleich der Sozialstruktur, politischen Einstellungen und Einstellungen zu Flüchtlingen zwischen AfD-WählerInnen und der WählerInnen der anderen Parteien. *Zeitschrift für vergleichende Politikwissenschaft* 12(1), 133–154.
- Harsanyi, J. C. (1977). *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*. Cambridge: Cambridge University Press.
- Hedström, P. (2005). *Dissecting the Social: On the Principles of Analytical Sociology*. Cambridge: Cambridge University Press.
- Hedström, P. and P. Bearman (2009). What is analytical sociology all about? An introductory essay. In P. Bearman and P. Hedström (Eds.), *The Oxford Handbook of Analytical Sociology*. Oxford: Oxford University Press.
- Hedström, P. and R. Swedberg (1996). Rational choice, empirical research, and the sociological tradition. *European Sociological Review* 12(2), 127–146.
- Hedström, P. and P. Ylikoski (2010). Causal mechanisms in the social sciences. *Annual Review of Sociology* 36, 49–67.
- Hedström, P. and P. Ylikoski (2014). Analytical sociology and rational choice theory. In G. Manzo (Ed.), *Analytical Sociology*, pp. 57–70. Chichester: Wiley.
- Hernes, G. (1992). We are smarter than we think: A rejoinder to Smelser. *Rationality and Society* 4(4), 421–436.
- Ho, T.-H., C. F. Camerer, and K. Weigelt (1998). Iterated dominance and iterated best response in experimental “p-beauty contests”. *American Economic Review* 88(4), 947–969.
- Hobbes, T. (2017 [1651]). *Leviathan*. Harmondsworth: Penguin.
- Hoem, J. M. (1991). To marry, just in case...: The Swedish widow’s-pension reform and the peak in marriages in December 1989. *Acta Sociologica* 34(2), 127–135.

- Homans, G. C. (1974). *Social Behavior: Its Elementary Forms*. Oxford: Harcourt Brace Jovanovich.
- Huber, J., J. W. Payne, and C. Puto (1982). Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis. *Journal of Consumer Research* 9(1), 90–98.
- Iannaccone, L. R. (1991). The consequences of religious market structure: Adam Smith and the economics of religion. *Rationality and Society* 3(2), 156–177.
- Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *American Economic Review* 93(5), 1449–1475.
- Kahneman, D. (2011). *Thinking, Fast and Slow*. New York: Farrar, Strauss and Giroux.
- Kahneman, D. and S. Frederick (2007). Frames and brains: Elicitation and control of response tendencies. *Trends in Cognitive Sciences* 11(2), 45–46.
- Kahneman, D. and A. Tversky (1979). Prospect theory: An analysis of decision under risk. *Econometrica* 47(2), 263–292.
- Kenrick, D. T. and V. Griskevicius (2013). *The Rational Animal: How Evolution Made Us Smarter Than We Think*. New York: Basic Books.
- Keren, G. and Y. Schul (2009). Two is not always better than one: A critical evaluation of two-system theories. *Perspectives on Psychological Science* 4(6), 533–550.
- Krottnerus, J. D. and T. N. Greenstein (1981). Status and performance characteristics in social interaction: A theory of status validation. *Social Psychology Quarterly* 44(4), 338–349.
- Kreps, D. M. and E. L. Porteus (1978). Temporal resolution of uncertainty and dynamic choice theory. *Econometrica* 46(1), 185–200.
- Kroneberg, C. (2005). Die Definition der Situation und die variable Rationalität der Akteure: Ein allgemeines Modell des Handelns. *Zeitschrift für Soziologie* 34(5), 344–363.

- Kroneberg, C. (2011). *Die Erklärung sozialen Handelns: Grundlagen und Anwendung einer integrativen Theorie*. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Kroneberg, C. (2014). Frames, scripts, and variable rationality: An integrative theory of action. In G. Manzo (Ed.), *Analytical Sociology*, pp. 95–123. Chichester: Wiley.
- Kroneberg, C. and S. Schulz (2018). Revisiting the role of self-control in situational action theory. *European Journal of Criminology* 15(1), 56–76.
- Kruglanski, A. W. and G. Gigerenzer (2011). Intuitive and deliberate judgments are based on common principles. *Psychological Review* 118(1), 97–109.
- Lengfeld, H. (2017). Die „Alternative für Deutschland“: Eine Partei für Modernisierungsverlierer? *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 69(2), 209–232.
- Leschziner, V. and G. Brett (2019). Beyond two minds: Cognitive, embodied, and evaluative processes in creativity. *Social Psychology Quarterly* 82(4), 340–366.
- Levine, D. K. (2012). *Is Behavioral Economics Doomed? The Ordinary Versus the Extraordinary*. Cambridge: Open Book Publishers.
- Levitt, S. D., J. A. List, and D. H. Reiley (2010). What happens in the field stays in the field: Exploring whether professionals play minimax in laboratory experiments. *Econometrica* 78(4), 1413–1434.
- Levitt, S. D., J. A. List, and S. E. Sadoff (2011). Checkmate: Exploring backward induction among chess players. *American Economic Review* 101(2), 975–990.
- Liberman, V., S. M. Samuels, and L. Ross (2004). The name of the game: Predictive power of reputations versus situational labels in determining prisoner’s dilemma game moves. *Personality and Social Psychology Bulletin* 30(9), 1175–1185.
- Liebe, U. and H. Beyer (2021). Examining discrimination in everyday life: A stated choice experiment on racism in the sharing economy. *Journal of Ethnic and Migration Studies* 47(9), 2065–2088.

- Lindenberg, S. (1992). The method of decreasing abstraction. In J. S. Coleman and T. J. Fararo (Eds.), *Rational Choice Theory. Advocacy and Critique*, pp. 4–20. Newbury Park: Sage Publications.
- Lindenberg, S. (1996). Constitutionalism versus relationalism: Two versions of rational choice sociology. In J. Clark (Ed.), *James S. Coleman*, pp. 200–212. New York: Routledge.
- List, J. A. (2011). Does market experience eliminate market anomalies? The case of exogenous market experience. *American Economic Review* 101(3), 313–317.
- Little, D. (1993). Evidence and objectivity in the social sciences. *Social Research* 69(2), 363–396.
- Lizardo, O. (2004). The cognitive origins of Bourdieu’s habitus. *Journal for the Theory of Social Behaviour* 34(4), 375–401.
- Lizardo, O., R. Mowry, B. Sepulvado, D. S. Stoltz, M. A. Taylor, J. van Ness, and M. Wood (2016). What are dual process models? Implications for cultural analysis in sociology. *Sociological Theory* 34(4), 287–310.
- Logan, J. A. (1996). Opportunity and choice in socially structured labor markets. *American Journal of Sociology* 102(1), 114–160.
- Loomes, G., C. Starmer, and R. Sugden (1991). Observing violations of transitivity by experimental methods. *Econometrica* 59(2), 425–439.
- Manzo, G. (2007). Progrès et «urgence» de la modélisation en sociologie. du concept de «modèle générateur» et de sa mise en œuvre. *L’Année sociologique* 57(1), 13–61.
- Manzo, G. (Ed.) (2021). *Research Handbook on Analytical Sociology*. Cheltenham: Edward Elgar.
- Manzo, G. (2022). *Agent-Based Models and Causal Inference*. Wiley series in computational and quantitative social sciences. Hoboken: John Wiley and Sons, Inc.

- May, K. O. (1954). Intransitivity, utility, and the aggregation of preference patterns. *Econometrica* 22(1), 1–13.
- Maynard Smith, J. (1982). *Evolution and the Theory of Games*. Cambridge: Cambridge University Press.
- McKelvey, R. D. and T. R. Palfrey (1992). An experimental study of the centipede game. *Econometrica* 60(4), 803–836.
- Mead, G. H. (2015 [1934]). *Mind, Self, and Society: The Definitive Edition*. Chicago: University of Chicago Press.
- Mesoudi, A. (2011). *Cultural Evolution: How Darwinian Theory Can Explain Human Culture and Synthesize the Social Sciences*. Chicago: University of Chicago Press.
- Miles, A., R. Charron-Chénier, and C. Schleifer (2019). Measuring automatic cognition: Advancing dual-process research in sociology. *American Sociological Review* 84(2), 308–333.
- Milgrom, P. (2004). *Putting Auction Theory to Work*. Cambridge: Cambridge University Press.
- Münch, R. (2007). *Soziologische Theorie: Band 2: Handlungstheorie*. Frankfurt a. M.: Campus.
- Murray, S. L., R. T. Pinkus, J. G. Holmes, B. Harris, S. Gomillion, M. Aloni, J. L. Derrick, and S. Leder (2011). Signaling when (and when not) to be cautious and self-protective: Impulsive and reflective trust in close relationships. *Journal of Personality and Social Psychology* 101(3), 485–502.
- Musgrave, A. (1981). ‘Unreal assumptions’ in economic theory: The F–twist untwisted. *Kyklos* 34(3), 377–387.
- Nagel, R. (1995). Unraveling in guessing games: An experimental study. *American Economic Review* 85(5), 1313–1326.

- Nagel, R. and F. F. Tang (1998). Experimental results on the centipede game in normal form: An investigation on learning. *Journal of Mathematical Psychology* 42(2), 356–384.
- Nakhaie, M. R., R. A. Silverman, and T. C. LaGrange (2000). Self-control and social control: An examination of gender, ethnicity, class and delinquency. *Canadian Journal of Sociology* 25(1), 35–59.
- Nash, J. F. (1950). Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences of the United States of America* 36(1), 48–49.
- Offerman, T., J. Sonnemans, and A. Schram (2001). Expectation formation in step-level public good games. *Economic Inquiry* 39(2), 250–269.
- Olinick, M. (2014). *Mathematical Modeling in the Social and Life Sciences*. Hoboken: John Wiley and Sons, Inc.
- Olson, M. (1965). *The Logic of Collective Action: Public Goods and the Theory of Groups*. Cambridge: Harvard University Press.
- Opp, K.-D. (1999). Contending conceptions of the theory of rational action. *Journal of Theoretical Politics* 11(2), 171–202.
- Opp, K.-D. (2011). Modeling micro-macro relationships: Problems and solutions. *Journal of Mathematical Sociology* 35(1-3), 209–234.
- Osborne, M. J. and A. Rubinstein (1998). Games with procedurally rational players. *American Economic Review* 88(4), 834–847.
- Osman, M. (2004). An evaluation of dual-process theories of reasoning. *Psychonomic Bulletin and Review* 11(6), 988–1010.
- Otsubo, H. and A. Rapoport (2008). Dynamic volunteer’s dilemmas over a finite horizon: An experimental study. *Journal of Conflict Resolution* 52(6), 961–984.
- Otte, G., T. Sawert, J. Brüderl, S. Kley, C. Kroneberg, and I. Rohlfling (2023). Gütekriterien in der Soziologie: Eine analytisch-empirische Perspektive. *Zeitschrift für Soziologie* 52(1), 26–49.

- Padilla-Walker, L. M. and G. Carlo (Eds.) (2014). *Prosocial Development: A Multidimensional Approach*. Oxford: Oxford University Press.
- Palacios-Huerta, I. and O. Volij (2009). Field centipedes. *American Economic Review* 99(4), 1619–1635.
- Pareto, V. (1935 [1917]). *The Mind and Society*. New York: Harcourt Brace Jovanovich.
- Parsons, T. (1937). *The Structure of Social Action: A Study in Social Theory With Special Reference to a Group of Recent European Writers*. New York: Free Press.
- Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology* 46(3), 598–609.
- Penner, L. A., B. A. Fritzsche, P. Craiger, and T. R. Freifeld (1995). Measuring the prosocial personality. In J. N. Butcher and C. D. Spielberg (Eds.), *Advances in Personality Assessment*, pp. 147–163. New York: Psychology Press.
- Popper, K. (2005 [1935]). *The Logic of Scientific Discovery*. New York: Routledge.
- Popper, K. (2013 [1994]). *All Life is Problem Solving*. London: Routledge.
- Rabin, M. (1998). Psychology and economics. *Journal of Economic Literature* 36(1), 11–46.
- Rand, D. G., J. D. Greene, and M. A. Nowak (2012). Spontaneous giving and calculated greed. *Nature* 489(7416), 427–430.
- Rand, D. G. and M. A. Nowak (2012). Evolutionary dynamics in finite populations can explain the full range of cooperative behaviors observed in the centipede game. *Journal of Theoretical Biology* 300, 212–221.
- Rapoport, A., W. E. Stein, J. E. Parco, and T. E. Nicholas (2003). Equilibrium play and adaptive learning in a three-person centipede game. *Games and Economic Behavior* 43(2), 239–265.

- Raub, W., V. Buskens, and M. A. L. M. van Assen (2011). Micro-macro links and microfoundations in sociology. *Journal of Mathematical Sociology* 35(1-3), 1–25.
- Raub, W., N. D. de Graafen, and K. Gërxhani (2022). Rigorous sociology. In K. Gërxhani, N. D. de Graafen, and W. Raub (Eds.), *Handbook of Sociological Science: Contributions to Rigorous Sociology*, pp. 2–19. Cheltenham: Edward Elgar.
- Ridgeway, C. L. and T. Kricheli-Katz (2013). Intersecting cultural beliefs in social relations: Gender, race, and class binds and freedoms. *Gender and Society* 27(3), 294–318.
- Rippl, S. and C. Seipel (2018). Modernisierungsverlierer, Cultural Backlash, Postdemokratie. *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 70(2), 237–254.
- Robinson, W. C. (1997). The economic theory of fertility over three decades. *Population studies* 51(1), 63–74.
- Rosenthal, R. W. (1979). Sequences of games with varying opponents. *Econometrica* 47(6), 1353–1366.
- Rubinstein, A. (1988). Similarity and decision-making under risk: Is there a utility theory resolution to the Allais paradox? *Journal of Economic Theory* 46(1), 145–153.
- Rubinstein, A. (1989). The electronic mail game: Strategic behavior under almost common knowledge. *American Economic Review* 79(3), 385–391.
- Rubinstein, A. (1998). *Modeling Bounded Rationality*. Cambridge and London: MIT Press.
- Rubinstein, A. (2006). *Lecture Notes in Microeconomic Theory*. Princeton: Princeton University Press.
- Rubinstein, A. (2007). Instinctive and cognitive reasoning: a study of response times. *Economic Journal* 117(523), 1243–1259.

- Rubinstein, A. (2013). Response time and decision making: An experimental study. *Judgment and Decision Making* 8(5), 540–551.
- Rubinstein, A. (2016). A typology of players: Between instinctive and contemplative. *Quarterly Journal of Economics* 131(2), 859–890.
- Rubinstein, A. and M. J. Osborne (2020). *Models in Microeconomic Theory*. Cambridge: Open Book Publishers.
- Schütz, A. (1944). The stranger: An essay in social psychology. *American Journal of Sociology* 49(6), 499–507.
- Schütz, A. (1990). *The Problem of Social Reality: Collected Papers 1* (6. ed.). The Hague: Nijhoff.
- Schwerter, F. and F. Zimmermann (2020). Determinants of trust: The role of personal experiences. *Games and Economic Behavior* 122, 413–425.
- Selten, R. (2002). What is bounded rationality? In G. Gigerenzer and R. Selten (Eds.), *Bounded Rationality: The Adaptive Toolbox*, pp. 13–36. Cambridge: MIT Press.
- Shariff, A. F. and A. Norenzayan (2007). God is watching you: Priming god concepts increases prosocial behavior in an anonymous economic game. *Psychological Science* 18(9), 803–809.
- Simon, H. A. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics* 69(1), 99–118.
- Simon, H. A. (1957). *Models of Man*. New York: Wiley.
- Simon, H. A. (1976). From substantive to procedural rationality. In T. J. Kastelein, S. K. Kuipers, W. A. Nijenhuis, and G. R. Wagenaar (Eds.), *25 Years of Economic Theory*, pp. 65–86. Boston: Springer.
- Simon, H. A. (1990). Bounded rationality. In J. Eatwell, M. Milgate, and P. Newman (Eds.), *The New Palgrave*, pp. 15–18. New York: W. W. Norton.

- Simpson, B. and H. A. Walker (2002). Status characteristics and performance expectations: A reformulation. *Sociological Theory* 20(1), 24–40.
- Simpson, B., R. Willer, and C. L. Ridgeway (2012). Status hierarchies and the organization of collective action. *Sociological Theory* 30(3), 149–166.
- Smelser, N. J. (1992). The rational choice perspective: A theoretical assessment. *Rationality and Society* 4(4), 381–410.
- Smith, E. R. and J. DeCoster (2000). Dual-process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems. *Personality and Social Psychology Review* 4(2), 108–131.
- Squire, L. R. (2004). Memory systems of the brain: A brief history and current perspective. *Neurobiology of Learning and Memory* 82(3), 171–177.
- Squire, L. R. and J. T. Wixted (2011). The cognitive neuroscience of human memory since H.M. *Annual Review of Neuroscience* 34, 259–288.
- Srull, T. K. and R. S. Wyer (1979). The role of category accessibility in the interpretation of information about persons: Some determinants and implications. *Journal of Personality and Social Psychology* 37(10), 1660–1672.
- Stanovich, K. E. (2011). *Rationality and the Reflective Mind*. New York: Oxford University Press.
- Stanovich, K. E. and R. F. West (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences* 23(5), 645–665.
- Stigler, G. J. and G. S. Becker (1977). De gustibus non est disputandum. *American Economic Review* 67(2), 76–90.
- Strack, F. and R. Deutsch (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review* 8(3), 220–247.
- Sumner, P. and M. Husain (2008). At the edge of consciousness: Automatic motor activation and voluntary control. *The Neuroscientist* 14(5), 474–486.

- Swidler, A. (1986). Culture in action: Symbols and strategies. *American Sociological Review* 51(2), 273–286.
- Thaler, R. H. (1980). Toward a positive theory of consumer choice. *Journal of Economic Behavior and Organization* 1(1), 39–60.
- Thomas, W. I. and D. S. Thomas (1928). *The Child in America: Behavior Problems and Programs*. New York: Knopf.
- Toplak, M. E., R. F. West, and K. E. Stanovich (2011). The cognitive reflection test as a predictor of performance on heuristics-and-biases tasks. *Memory and Cognition* 39(7), 1275–1289.
- Toplak, M. E., R. F. West, and K. E. Stanovich (2014). Assessing miserly information processing: An expansion of the cognitive reflection test. *Thinking and Reasoning* 20(2), 147–168.
- Tutić, A. (2014). Procedurally rational volunteers. *Journal of Mathematical Sociology* 38(3), 219–232.
- Tutić, A. (2015a). Revealed norm obedience. *Social Choice and Welfare* 44(2), 301–318.
- Tutić, A. (2015b). Warum denn eigentlich nicht? Zur Axiomatisierung soziologischer Handlungstheorie. *Zeitschrift für Soziologie* 44(2), 83–98.
- Tutić, A. (2022). Cultural orientations and their influence on social behaviour: Catalysation and suppression. *Journal for the Theory of Social Behaviour* 52(3), 438–453.
- Tutić, A. and S. Grehl (2017). A note on disbelief in others regarding backward induction. *Games* 8(3), 33.
- Tutić, A. and S. Grehl (2018). Status characteristics and the provision of public goods: Experimental evidence. *Sociological Science* 5, 1–20.
- Tutić, A. and S. Grehl (2021). Implizite Einstellungen, explizite Einstellungen und die Affinität zur AfD. *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 73(3), 389–417.

- Tversky, A. (1969). Intransitivity of preferences. *Psychological Review* 76(1), 31–48.
- Tversky, A. (1972). Elimination by aspects: A theory of choice. *Psychological Review* 79(4), 281–299.
- Tversky, A. and D. Kahneman (1974). Judgment under uncertainty: Heuristics and biases. *Science* 185(4157), 1124–1131.
- Tversky, A. and D. Kahneman (1981). The framing of decisions and the psychology of choice. *Science* 211(4481), 453–458.
- Udehn, L. (2002). *Methodological Individualism: Background, History and Meaning*. London: Routledge.
- Vaisey, S. (2009). Motivation and justification: A dual-process model of culture in action. *American Journal of Sociology* 114(6), 1675–1715.
- Vaisey, S. and O. Lizardo (2010). Can cultural worldviews influence network composition? *Social Forces* 88(4), 1595–1618.
- van Bavel, J. J., Y. Jenny Xiao, and W. A. Cunningham (2012). Evaluation is a dynamic process: Moving beyond dual system models. *Social and Personality Psychology Compass* 6(6), 438–454.
- Vila-Henninger, L. A. (2015). Toward defining the causal role of consciousness: Using models of memory and moral judgment from cognitive neuroscience to expand the sociological dual-process model. *Journal for the Theory of Social Behaviour* 45(2), 238–260.
- Vila-Henninger, L. A. (2021). A dual-process model of economic behavior: Using culture and cognition, economic sociology, behavioral economics, and neuroscience to reconcile moral and self-interested economic action. *Sociological Forum* 36(S1).
- von Neumann, J. and O. Morgenstern (1944). *Theory of Games and Economic Behavior*. Princeton: Princeton University Press.

- Walder, A. G. (1992). Property rights and stratification in socialist redistributive economies. *American Sociological Review* 57(4), 524–539.
- Weber, M. (1978). *Max Weber: Selections in Translation*. Cambridge: Cambridge University Press.
- Weber, M. (1978 [1921/1922]). *Economy and Society*. Berkley and Los Angeles: University of California Press.
- Weber, R. A. (2001). Behavior and learning in the “dirty faces” game. *Experimental Economics* 4 (3), 229–242.
- Webster, Jr., M. and J. E. Driskell, Jr. (1978). Status generalization: A review and some new data. *American Sociological Review* 43(2), 220–236.
- Weesie, J. (1993). Asymmetry and timing in the volunteer’s dilemma. *Journal of Conflict Resolution* 37(3), 569–590.
- Whitney, P., C. A. Rinehart, and J. M. Hinson (2008). Framing effects under cognitive load: The role of working memory in risky decisions. *Psychonomic Bulletin and Review* 15(6), 1179–1184.
- Wilson, T. D., S. Lindsey, and T. Y. Schooler (2000). A model of dual attitudes. *Psychological Review* 107(1), 101–126.
- Wippler, R. and S. Lindenberg (1987). Collective phenomena and rational choice. In J. C. Alexander, B. Giesen, R. Münch, and N. J. Smelser (Eds.), *The Micro-Macro Link*, pp. 135–152. Berkley and Los Angeles: University of California Press.
- Ylikoski, P. (2021). Understanding the Coleman boat. In G. Manzo (Ed.), *Research Handbook on Analytical Sociology*, pp. 49–63. Cheltenham: Edward Elgar.
- Zahle, J. (2014). Holism, emergence, and the crucial distinction. In J. Zahle and F. Collin (Eds.), *Rethinking the Individualism-Holism Debate*. Heidelberg: Springer.

-
- Zahle, J. and F. Collin (Eds.) (2014). *Rethinking the Individualism-Holism Debate: Essays in the Philosophy of Social Science*. Heidelberg: Springer.
- Zaki, J. and J. P. Mitchell (2013). Intuitive prosociality. *Current Directions in Psychological Science* 22(6), 466–470.
- Zelditch, Jr., M., P. Lauderdale, and S. Stublarec (1980). How are inconsistencies between status and ability resolved? *Social Forces* 58(4), 1025–1043.

Articles

Verhaltensökonomik und Begrenzte Rationalität (Behavioral economics and bounded rationality)

Sascha Grehl.

in: Tutić, Andreas (Ed.): Rational Choice. Berlin: De Gruyter Oldenbourg, 150–178. Translated from German.

Experimental Evidence on Iterated Reasoning in Games

Sascha Grehl, Andreas Tutić

PLoS ONE 10(8), e0136524 (2015).

RESEARCH ARTICLE

Experimental Evidence on Iterated Reasoning in Games

Sascha Grehl*, Andreas Tutić

Institute of Sociology, Leipzig University, Leipzig, Germany

* sascha.grehl@uni-leipzig.de



CrossMark
click for updates

 OPEN ACCESS

Citation: Grehl S, Tutić A (2015) Experimental Evidence on Iterated Reasoning in Games. PLoS ONE 10(8): e0136524. doi:10.1371/journal.pone.0136524

Editor: MariaPaz Espinosa, University of the Basque Country, SPAIN

Received: January 15, 2015

Accepted: August 5, 2015

Published: August 27, 2015

Copyright: © 2015 Grehl, Tutić. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by Deutsche Forschungsgemeinschaft, grant number: TU 409/1-1, receiver of the funding: Andreas Tutić, website: <http://www.dfg.de>. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We acknowledge support from the German Research Foundation (DFG) and Universität Leipzig within the program of Open Access Publishing.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

We present experimental evidence on two forms of iterated reasoning in games, i.e. backward induction and interactive knowledge. Besides reliable estimates of the cognitive skills of the subjects, our design allows us to disentangle two possible explanations for the observed limits in performed iterated reasoning: Restrictions in subjects' cognitive abilities and their beliefs concerning the rationality of co-players. In comparison to previous literature, our estimates regarding subjects' skills in iterated reasoning are quite pessimistic. Also, we find that beliefs concerning the rationality of co-players are completely irrelevant in explaining the observed limited amount of iterated reasoning in the dirty faces game. In addition, it is demonstrated that skills in backward induction are a solid predictor for skills in iterated knowledge, which points to some generalized ability of the subjects in iterated reasoning.

1 Introduction

Recently the question of how humans actually reason in game-theoretical problems has received some attention in the literature [1–3]. Experimental evidence as well as casual introspection suggest that orthodox decision and game theory needs fundamental modifications to bolster its explanative and predictive potential regarding human behavior. In recent years scholars have worked on both providing new models of (interactive) decision making [4–6] as well as identifying the main properties of standard theory which undermine its explanative power [7–9].

Our paper contributes to the second, empirical branch of this literature. Our research focuses on subjects' ability to perform iterated reasoning, their belief about how well others might do this, and how the subjects are influenced by this belief. We concentrate on iterated reasoning, because many game-theoretical solution concepts such as iterated dominance or backward induction explicitly require that subjects perform at least several steps of iterated reasoning. That is, iteration is directly involved in the definitions of these solution concepts. In addition, even seemingly innocuous solution concepts, such as the Nash equilibrium, in fact involve iterated reasoning in the form of 'common knowledge' regarding various aspects of the game, as revealed by epistemic game theory [10]. Hence, iterated reasoning plays a major role

in noncooperative game theory, and the assessment of whether humans are actually capable of engaging in iterated reasoning is of great interest.

We are of course not the first to study iterated reasoning in games (see next section). However, our experimental study improves upon the existing literature in the following key aspects. First, previous studies often suffer from the problem that the ability to engage in iterated reasoning and the belief in the ability of one's co-players cannot be separated properly. That is, if some particular subject shows a limited amount of iterated reasoning, then it is impossible to tell whether this limitedness is due to a bounded ability of the subject or her beliefs regarding the behavior of her co-players. Second, we measure two different forms of iterated reasoning, i.e. backward induction and interactive knowledge, and hence explore whether there is an underlying propensity of subjects to engage in iterated reasoning or if its practice is more specific to concrete decision problems. Third, for each form of iterated reasoning under consideration we obtain multiple observations from each subject while for the most part excluding learning effects. Hence, our measure of the depth of iterated reasoning is in a sense more reliable than typically found in the literature.

The rest of the paper is organized as follows. Section 2 provides a small review of the existing experimental literature on iterated reasoning. Section 3 provides the details on our experimental design. In section 4 we present our results, the fifth section concludes.

2 Related literature

Stahl and Wilson [11, 12] as well as Nagel [13] pioneered the systematic study of varying degrees of rationality via experiments. They formulated the so-called level- k model which rests upon the idea that players have varying depths of iterated reasoning. In this model, each subject is characterized by a certain level $k \in \mathbb{N}$, which denotes how many steps of iterated best responses the subjects apply to their belief of what a level-0 player would do. A level-0 type ($L0$) is defined as non-strategic, which means that she has no particular belief about the strategies of others and therefore follows a salient decision rule. A level-1 type ($L1$) believes that all other players are $L0$ types and hence best-responds to this $L0$ strategy. In general a level- k type (Lk), for any $k \geq 1$, best-responds to the belief that all others are at maximum level- $(k - 1)$ types. Depending on assumptions with respect to the player's expectations regarding the distribution of her co-players' types [14, 15] and specification of the $L0$ strategy [16, 17], data on observed behavior in games can be utilized to assign subjects to types (i.e. levels). Generally speaking, in experimental studies using the level- k approach levels greater than 3 are rarely observed [18]. Clearly, the level k -approach involves iterated reasoning, because higher-level players need to 'calculate' all the choices of lower-level players to determine their own choice.

We now turn to the experimental literature that more explicitly refers to iterated reasoning. Two main threads can be identified here. The first one focuses on how many steps of iterated reasoning are performed in general by humans, while the second one pays attention to the process in which subjects can learn to engage in iterated reasoning [19]. Put briefly, it is—as shown by the level- k literature—observed that subjects seldom use more than 3 steps of iterated reasoning. With respect to learning, these studies find that both repetition [2, 20] as well as time for reflection [21] considerably improve subjects' performance on iterated reasoning. Recently, in both strands the influence of general cognitive skills, such as short-term memory capacity or IQ, has also moved increasingly into focus. It turns out that higher cognitive skills positively affect the iterated reasoning performance [1, 22, 23], whereas shocking these skills decreases the performance [23].

Finally, we mention a number other contributions to the literature which relate to our experiment but do not necessarily belong to some identifiable strand. Grosskopf and Nagel [8],

Agranov et al. [24], and Carpenter et al. [23] study the question of whether subjects are actually able to find best responses in situations in which the beliefs are either irrelevant (i.e. a weakly dominant strategy exists), exogenously imposed, or measured ex post. Somewhat depressingly, it turns out that many subjects fail to give best responses, even in not-too-demanding decision situations (e.g. two-person beauty contests). Rubinstein [25, 26] advocates the study of response times; he shows over a wide variety of games and decision situations that patterns in response times might plausibly be related to different types of decision procedures and heuristics.

3 Experimental design and methods

This section contains all information on the methodological aspects of our study. First, we motivate our experiment against the background of the reviewed literature. Second, we describe our measurement instruments, i.e. the cognitive reflection test, the hit game, and the dirty faces game. Third, we provide our experimental design and outline the course of a session.

3.1 Motivation

A common problem in the level- k approach as well as in the literature on iterated reasoning is the fact that experimental designs do not allow any differentiation between two possible explanations for the observed limited amount of iterated reasoning [27]. That is, limits in the performed depth of iterated reasoning can be explained by limited cognitive abilities or by beliefs regarding the amount of iterated reasoning performed by the co-players. Note that the level- k approach does not necessarily commit itself with respect to the question of whether the types refer to cognitive abilities regarding the depths of iterated reasoning or to the beliefs regarding the behavior of the co-players. An observed level- k player might actually be capable of determining many higher levels of best responses, but abstains from doing so because she believes that her co-players are rather unsophisticated. With respect to the literature on iterated reasoning, a similar problem arises because out-of-equilibrium behavior can always be justified by a lack of rationality or by a lack of belief in the rationality of the co-players, as long as no weakly dominant strategies are involved. Hence, the disentanglement of cognitive ability to and belief in iterated reasoning is a natural next step in the empirical study of human behavior in strategic situations.

Surprisingly, the literature on iterated reasoning has made only little effort to assess whether humans have some kind of general capacity to engage in iterated reasoning or the performance crucially depends on the specific cognitive task. Generally, studies focus on one form of iterated reasoning, i.e. either on iterated dominance (beauty contest), backward induction (centipede game and hit game), or interactive knowledge (dirty faces game). This way it was clearly demonstrated that there is a considerable amount of heterogeneity in the displayed depth of iterated reasoning between subjects. However, the question of whether there is heterogeneity within subjects, i.e. whether a subject's performance in iterated reasoning depends on the form of reasoning, has not been pursued. Many studies only measure one form of iterated reasoning. Even if several forms of iterated reasoning are involved [1, 23], scholars have not pursued the question of the relationship between them.

Finally, in most studies that are not interested in learning effects, only a very small amount of observations are used to assess the performance of individual subjects in interactive reasoning. However, this practice is prone to produce unreliable measures, since subjects can perform well simply due to chance. Hence, it is important to collect multiple measurements and control for learning effects at the same time.

In our study, we want to deal with all of these concerns. That is, we control for the beliefs of the subjects by either using games in which beliefs do not matter or providing the subjects with an exogenous belief. The latter is achieved by substituting the co-players with an algorithm programmed to play perfectly rationally and communicating this to the subjects. Further, we measure subjects' performance on two forms of iterated reasoning, i.e. backward induction and interactive knowledge. For this purpose, we use the hit game and the dirty faces game (see section 3.2). Lastly, for the sake of more reliable measurement, subjects had to play these games several times. In contrast to learning studies, we try to handicap learning as much as possible by employing a number of countermeasures (see section 3.3).

3.2 Measurement instruments

As is common in the literature on iterated reasoning [1, 22], we tried to elicit some cognitive capabilities of our subjects in addition to their behavior in games. For this purpose we picked the cognitive reflection test (CRT) introduced by Frederick [28]. The CRT consists of three questions (see Table 1), which "are 'easy' in the sense that their solution is easily understood when explained, yet reaching the correct answer often requires the suppression of an erroneous answer that springs 'impulsively' to mind" [28]. We chose the CRT because it seems to be related to the idea of Kahneman [29] and many other researchers [30] that humans have two systems of thought for solving problems, i.e. a spontaneous and hence barely conscious one as well as a slow but more reflective one, which we find intriguing.

To measure iterated reasoning in the form of backward induction, we use the hit game [2, 20, 23]. The hit game is defined by a number $m \in \mathbb{N}$ and an interval $[a, b]$ with $0 < a < b$ and $a, b \in \mathbb{N}$. Two players alternately pick an integer from $[a, b]$. These numbers are added up. The player who reaches m or surpasses it wins the game. Since the hit game is a sequential game with complete and perfect information backward induction can be applied to determine the subgame perfect equilibrium: Depending on the game parameters either the first or the second player can ensure a win by consistently forcing the other player into so-called losing positions while the other player is incapable of influencing this outcome (For more details on the solution, see appendix). The hit game provides a straightforward and easy to interpret measure of iterated reasoning; the depth of reasoning required to solve this game can simply be equated with the number of picks of the winning player on the backward induction path. Interestingly, the level- k approach somehow fails in this game. Since no specific salient strategy exist, common practice suggests identifying $L0$ with the uniform distribution on the players' respective strategy spaces [11]. However, at the start of the game one player is in a position to be able to force a win. This player has a weakly dominant strategy, hence already $L1$ -players need to apply backward induction perfectly. As a consequence, concerning the player in the winning position, the level- k approach can only discriminate between $L0$ - and $L1$ -types; this fails to capture the intuition that players might very well be able to solve 'small' hit games, but fail in hit games of considerable complexity.

Table 1. The cognitive reflection test (CRT).

CRT1	A bat and a ball cost EUR 1.10 in total. The bat costs EUR 1.00 more than the ball. How much does the ball cost?
CRT2	If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets?
CRT3	In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake?

doi:10.1371/journal.pone.0136524.t001

We now turn to the dirty faces game, which is commonly used to measure iterated reasoning in the form of interactive knowledge [19, 31, 32]. Each subject is assigned a type, either X or O. Each player knows the types of all others but not her own. However, it is publicly announced that at least one player is an X-type. Then the game proceeds in turns, with subjects privately choosing one of the three possible announcements 'I am an X-type' (X), 'I am an O-type' (O), or 'I don't know my type' (U). When everyone has chosen an announcement, these are made public and a new turn begins. The aim of the game is to logically deduce one's own type and to publicly announce it as quickly as possible. Under the condition of common knowledge of rationality, standard arguments from interactive epistemology suggest the following solution [19]: Suppose there are k X-types. Everybody announces U in turn 1, ..., $k - 1$. In turn k all X-types announce X, while O-types continue to announce U. In turn $k + 1$ the O-types announce O (For more details on the solution, see appendix). Finally, we note that the depth of reasoning required to solve the dirty faces game differs between X- and O-types, because the O-types deduce their type on the basis of the observed announcements of X-types. For both types we can simply equate the depths of reasoning involved in solving the game with the number of (subjectively) observed X-types + 1 (Adding 1 simply normalizes our measure. That is, in a dirty faces game with exactly one X-player, our measure gives 1 for this X-player and 2 for each of the O-players).

3.3 Subject pool, procedure, and design

The experiment was conducted in the experimental laboratory of Leipzig University from fall 2013 through summer 2014. The experiment was conducted in accordance with the Declaration of Helsinki and all procedures were approved by the Institute of Sociology of the Leipzig University. Participants were recruited via the internet recruitment tool hroot [33]. Information about the duration, the payment, and the confidentiality was provided to participants prior to signing up for the experiments. By voluntarily signing up for the experiments via our website, participants provided written consent to participate in the study. Each participant received a EUR 5.00 show-up-fee and could earn additional money during the course of the experiment. In total 269 subjects attended, earning an average of EUR 14.55 in 17 sessions lasting about 1.5 hours.

To begin with, participants received written instructions with general information about the experiment, the fact that communication was prohibited, payment, anonymity, and time restrictions. We pointed out that for each question and decision there would be a guideline time, which could be exceeded but should not be exceeded constantly. Then participants had the opportunity to ask questions about these general rules.

After that the CRT was conducted. The whole experiment was programmed and conducted with the software z-Tree [34]. Participants had 2 minutes for each question. Participants received no feedback between the questions, but at the end of the test the number of correct answers was displayed; these were rewarded with EUR 0.50 each.

Thereafter the instructions for the hit game were displayed on the screen. Participants encountered a practice round (hit0) followed by 7 distinct rounds (see Table 2), which were rewarded monetarily with EUR 0.50 each if won. Concerning the interval $[a, b]$ from which the subjects had to pick their numbers, b was fixed at 3, while a alternated between 1 and 2, so as to reduce learning effects. Further m increases over time, raising the complexity of the hit games. The complexity of a game refers to the number of correct decisions necessary to win a certain game, i.e. the depth of iterated reasoning required to solve them. Rather than letting the participants play against each other, they played against an algorithm, which was programmed to play the game rationally, i.e. as long as the algorithm was in the losing position it adds the

Table 2. The hit games implemented.

	hit0	hit1	hit2	hit3	hit4	hit5	hit6	hit7
Starting value (<i>m</i>)	6	6	11	11	13	13	18	18
Minimum (<i>a</i>)	1	2	1	2	1	2	1	2
Complexity (<i>C</i>)	2	2	3	3	4	3	5	4

doi:10.1371/journal.pone.0136524.t002

smallest number, which minimizes the chance for the subject to stay on the winning path by guessing. Once the subject made a mistake the algorithm stayed on the winning path until it won the game. This fact was communicated to the subjects in advance. We used algorithms for two reasons: First, this way all subjects could start each round in the same (winning) positions, resulting in more interpretable observations, and second, we wanted to rule out possible effects of other-regarding preferences.

When all subjects had finished the hit games, we handed out the written instructions for the dirty faces game. Participants then had about 15 minutes to study these, to ask questions, and to complete a quiz concerning the payoff structure and game mechanics. Irrespective of whether the subject answered a question right or wrong, after each question the correct answer and an explanation were provided. If no further questions were asked, we proceeded with the experiment.

To disentangle the influence of ability to engage in iterated reasoning and the beliefs in the iterated reasoning of the co-players, the subjects played both with human co-players (HU version) and with an algorithm (AI version). To control for learning effects, we implemented two experimental treatments: In the AH treatment, the subjects first played with the algorithm and then with fellow subjects, while the order was reversed in the HA treatment.

At the beginning of the AI version subjects were informed that they were playing with an algorithm, which had been programmed to logically deduce its type correctly from the given information and to assume that the subject will do the same. Since subjects were informed about this, they could rely upon the rationality of their (algorithmic) co-players. The observed performance in the AI version could thus be attributed directly to the ability to engage in this kind of reasoning. To rule out subject confusion [27] we stopped a game in the AI version whenever an illogical announcement from the subject would imply that she would observe the algorithm announcing a type which contradicted its true type. This can happen for example when the subject is the only X-type but irrationally announces U in the first turn. The algorithm would then logically correct but de facto wrongly infer that it must be an X-type. In such cases the game stopped after turn 2.

During each new round subjects were randomly and anonymously paired in groups. We varied the group size in both versions across sessions from four to six, to check whether there were any effects of group size. During a session group size was held constant, hence 112 subjects played in groups of four, 85 in groups of five, 72 in groups of six. To guarantee better comparability between the games of different group sizes, we restricted all games to the seven possible situations someone in a four-person group could be confronted with. These are the trivial situation where the player observes no X-type plus the constellations where the player observes one to three X-types while she is either an X-type herself or not. Further we balance the constellations for the subjects in such a way that they encounter each problem besides the trivial one at least once in both versions of the dirty faces game. Remember that the complexity of a certain dirty faces game is equal to the number of observed X-types + 1. In the AI version subjects simply played all seven constellations, resulting in 269 observations per constellation.

Table 3. The df games implemented and observations in the AI/HU version.

Player's type	Complexity (C)				
	1	2	3	4	5
X	269/111	269/286	269/378	269/332	-/-
O	-/-	269/409	269/409	269/232	-/76*

*only in HU version with group size of 5 or 6

doi:10.1371/journal.pone.0136524.t003

In the HU version subjects played, depending on group size, 7 to 10 games. Here another constellation could be observed if group size exceeded four, that is the situation where an O-type observes four X-types (see Table 3).

A round ended for a subject as soon as she announced a type or, alternatively, after the sixth turn (theoretically, 6 turns would suffice to solve any constellation of the dirty faces game used in our study). Each player who correctly announced her type got a lottery ticket, which provided the opportunity to win EUR 0.50. If the subject announced the wrong type, she got nothing. The probability that a ticket would win was 100% minus 5% for each time the receiver of the ticket had chosen U in the respective round. This way we guaranteed that announcing one's type as soon as it was possible on logical grounds was an attractive strategy for maximizing monetary payoffs. To ensure, further, that guessing was an unattractive strategy, players choosing U until the end of the round got a 60% ticket. Such a ticket was better than guessing in the first or any subsequent turn, assuming a 50% probability of guessing correctly. We use lottery tickets instead of direct payoffs due to the fact that in theory individual risk preferences should be ruled out by this procedure. Further, we opted for an individual rewarding based on the actual type (in contrast to making payment conditional on the whole group determining their type correctly [35]) because otherwise we would lose the strategic uncertainty and hence makes beliefs again completely irrelevant in the HU version. To reduce learning effects, no feedback was given between the rounds.

In the last stage of the experiment, the subjects filled in a short questionnaire. They were asked to specify their gender, age, field of study, and previous knowledge of game theory or logic.

4 Results

The sample for the entire study was 269 subjects, leading to 1883 observations for the hit game and the dirty faces games (henceforth, df games) in the AI version, as well as to 2233 observations for the df game in the HU version on the level of individual decision making. Nearly two thirds (63%) of our subjects were women and almost all were students (93%). The mean age was 24 years. We asked subjects about their field of study and if they had ever read a book (14%) or attended a course (25%) on game theory or logic. However, besides gender none of these variables turned out to be related to iterated reasoning in any conceivable way.

4.1 Cognitive reflection test

The numbers of correctly solved CRT questions (CRT score) are relatively evenly distributed, ranging from 21 to 30% (see Fig 1 (a)), which results in a mean score of 1.45 correct answers. Males score on average 0.38 points higher ($p = 0.005$; t-test) and this effect pervades through all three questions. Fig 1 (b) reveals that CRT2 and CRT3 are easier for our subjects. The internal

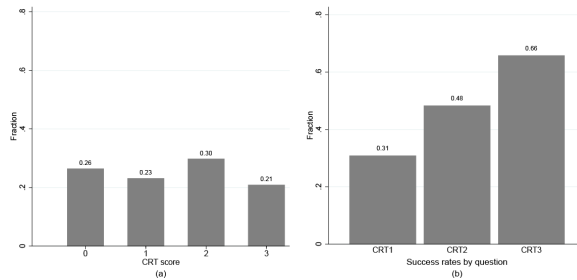


Fig 1. Subjects' performance in the cognition reflection test.

doi:10.1371/journal.pone.0136524.g001

consistency of the construct is acceptable (Cronbach's $\alpha = 0.63$). Hence, in the following we use the CRT score as a measure for the non-impulsiveness of thinking.

4.2 Hit game

In panel (a) of Fig 2 one can see the distribution of correctly solved hit games (hit score). Over two thirds of our subjects could solve only 2 or 3 problems, resulting in a mean of 2.58 solved games and immediately refuting the idea of perfectly rational actors. Panel (b) provides a compact overview about the observed behavior of our subjects in hit games. To interpret this graph, we introduce the notion of 'type of error'. An error ℓ refers to a wrong choice in a situation in which ℓ correct choices from the subject are still required to win the game. The graph shows the distribution of errors for each hit game under consideration. For example, consider hit7. To win this game, the subject needs to make four consecutive choices and all of these choices have to be correct. We see that approximately half of subjects fail in the very first decision, i.e. they make an error 4. From the subjects who do not make an error 4, approximately one third make an error of type 3. Then, a small minority of subjects make an error 2. Finally, errors of type 1 are absent in hit7, such that approximately 30% of subjects win the game.

The graph reveals several important patterns in the observed behavior. First of all, error 1 occurs virtually never (only once in hit2). Since an error of type 1 refers to a trivial situation in which the subjects were in a position to win the game directly, this shows that subjects had

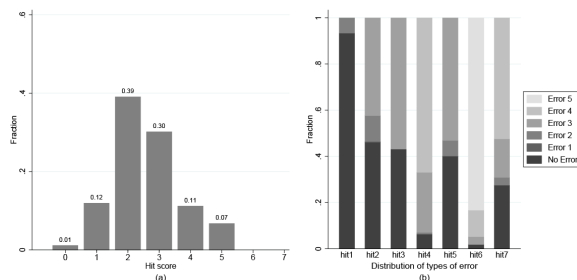


Fig 2. Subjects' performance in the hit games.

doi:10.1371/journal.pone.0136524.g002

Table 4. The *i*-index for the hit games.

<i>i</i> -index	Fraction
1	.07
2	.87
3	.06

doi:10.1371/journal.pone.0136524.t004

understood the rules of the game and were motivated to win. More interestingly, the fact that errors of type 2 occurred only rarely means that most subjects are reliably capable of anticipating one action of their co-player. However, the graph also shows that subjects did not substantially outperform random guessing in situations involving the anticipation of two or more actions by the co-player. Consider errors of type 3. Recall that in hit1, hit3, hit5, and hit7 the action space contained only 2 and 3, while in the even hit games the action space contained 1, 2, and 3. Hence, random guessing implies error fractions of 50% for odd hit games and approximately 33% for even hit games. We see that in hit2 and hit7 subjects make less errors of type 3 than predicted by random guessing. In all remaining games, the rate of errors of type 3 roughly equals the rate predicted by random guessing. Hence, there seems to be a small minority of subjects who are able to anticipate two actions of the co-player. Errors of type 4 occur more or less exactly in the proportion predicted by guessing. Surprisingly, errors of type 5 occur more frequently than predicted by random guessing.

A main advantage from our design is that multiple observations on the hit game allow a more reliable measurement of the depths of iterated reasoning in the form of backward induction. To separate success due to guessing from iterated reasoning, we construct an index in which a subject gets the index *i* if she was able to solve at least all games with 1, . . . , *i* iteration steps involved, but fails at a game with *i* + 1 iteration steps. Table 4 shows the result. Almost every participant could solve problems involving two or less steps of backward induction, but only a very small minority of 6% were able to reliably solve problems involving three steps of iteration. Finally, nobody among our 269 subjects was able to perform four or more steps reliably.

Note that our estimates are quite pessimistic in comparison to the literature [13, 36]. We provide two additional sources of evidence to back up our estimates. First, we calculate the expected frequencies of solved hit games based on the estimated distribution of the *i*-index and the assumption that subjects guess randomly in situations that involve more steps of iterations than indexed by their respective *i*-index. The result is shown in panel (a) of Fig 3. Surprisingly,

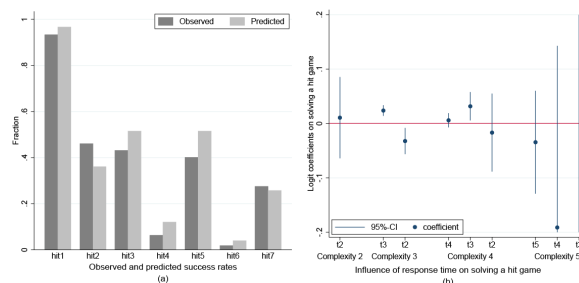


Fig 3. Success rates and response time analysis for the hit games.

doi:10.1371/journal.pone.0136524.g003

our estimates of the *i*-index combined with the assumption of random guessing in overcomplex situations tend to overestimate the proportion of successful subjects. We conclude that our estimates of the *i*-index do not underestimate the abilities of our subjects for iterated reasoning, but probably overestimate these abilities.

A second source of validation of our estimates comes from an analysis of response times (For the analysis we used pure response times. The analysis was conducted additionally with log response times, but this did not alter the results substantially). Panel (b) in Fig 3 plots the influence of response times on the probability of correctly solving a hit game from logistic regressions which control for gender, CRT score, and size of action space (Fully specified models are provided in Table 5. To achieve a favorable scale the CI is truncated at .2 and -.2, respectively). These regressions were run separately for each decision time grouped by category of complexity. Because each time subjects successfully stayed on the backward induction path the complexity of a game is reduced by one, the different decision times are labeled with a number referring to current complexity of the problem. Note that in each category we dropped the last decision time t1, i.e. the time when a direct win was possible, because, as mentioned earlier, nearly every subject solved this problem, leading to a lack of variance in the dependent variable. In addition, we had to drop t2 for games of complexity 5 for the same reason. Also, note that the estimates of the effects of the response times of decisions are based only on those subjects who did not fail in previous decisions in the respective game.

The graph reveals that the amount of time subjects take to think about overcomplex problems, i.e. the first decision in games with complexity 4 (t4) as well as the first two decisions in games with complexity 5 (t5, t4), doesn't matter with respect to the probability of solving the games. That is, these problems are too elaborate for the subjects and hence it doesn't matter how much time they invest. In stark contrast, the first decision matters for games of complexity

Table 5. Logit regressions on solving a hit game.

<i>hit game solved</i>	model			
	(1)	(2)	(3)	(4)
Complexity				
2	2.804**	3.004**	3.016**	3.031**
3 (reference)				
4	-1.267**	-0.948**	-0.947**	-0.957**
5	-3.469**	-2.754**	-2.747**	-2.777**
Size of action space	-0.329**	-0.328*	-0.328*	-0.331*
Response Time				
t2 (sec)		0.019	0.018	0.017
t3 (sec)		0.025**	0.024**	0.023**
t4 (sec)		0.007	0.006	0.005
t5 (sec)		-0.020	-0.022	-0.022
Gender (1 = male)			0.278*	0.211
CRT score				0.198**
Constant	-0.827**	-1.265**	-1.351**	-1.601**
Observations	1883	1883	1883	1883
Pseudo R ²	0.287	0.299	0.301	0.307

* *p* < 0.05

** *p* < 0.01

doi:10.1371/journal.pone.0136524.t005

3 (t3), as does the second decision for games of complexity 4 (t3). This conforms to our aforementioned finding that a considerable portion of our subjects are cognitively able to perform the required steps of reasoning in at least some of these games. Additionally, note that decisions in (sub-)games of complexity 2 are not affected by the amount of time invested in making the decisions (In games of complexity 3 we even found a negative influence of response time 2 (t2)). This finding makes sense, because the application of backward induction involves determining each choice right at the start of the (sub-)game which is simple enough to be solved by a considerable portion of the subjects, i.e. (sub-)games of complexity 3. Finally, the fact that the response time of the third decision (t3) in the complexity 5 category does not affect the probability of solving this game, which should be expected according to our reasoning, is most likely due to the small fraction of players who were actually lucky in their first two guesses such that they still had a chance of winning the game (only 14 subjects).

Against this background, we feel that our assessment of our subjects' skills in iterated reasoning is very solid indeed. We now turn to some other interesting aspects of the observed behavior in hit games. First, note that the hit games used in our study are structurally related in various ways. Some of the simpler games are 'contained' in more complex games, which should facilitate the application of backward induction. For example, the reasoning involved in solving hit3 is useful for solving hit5. That is, provided that the subject succeeded in hit3, she knows that the co-player has a winning strategy if she picks 2 in her first choice in hit5. Hence, she knows that it cannot be a bad idea to pick 3, the only alternative to 2 in this game. Similar relationships hold between hit1 and hit3, hit2 and hit4, hit4 and hit6, hit2 and hit6 as well as between hit5 and hit7. We also implemented 'traps'. Hit2 and hit3 both have $m = 11$, but the minimal pick equals 1 in hit2 and 2 in hit3. Consequently, backward induction dictates picking 3 in hit2 and 2 in hit3 as the respective first decision. The pairs hit4 and hit5 as well as hit6 and hit7 are traps too.

Indeed we find evidence for both backward induction as well as for sloppy, short-cutting reasoning. Table 6 contains the results of χ^2 based measures of association. First, note that all of the significant associations are descriptively positive. In three instances we find strong associations. The fact that the hit1 and hit3 are not related is not too troubling, considering that there is almost no variance in hit1. More puzzling is the finding that solving hit2 does not facilitate the solutions of hit4 and hit6. We speculate that this is due to our 'trap', which might somehow undermine the faith of our subject in the use of their solution of hit2 for more complex hit games. In fact the first trap worked fine; we find a strong negative association between hit2 and hit3, i.e. $\chi^2 = 33.606$, $p < 0.001$ and $\phi = -0.354$. The other two traps did not work, most likely because the subjects had learned their lesson.

Table 5 concludes our analysis of the hit game. It contains the results of three logistic regressions which estimate the probability of solving a hit game as a function of parameters of the hit

Table 6. Pearson's chi-squared test for independence.

hit games	χ^2	p	ϕ
hit1—hit3	0.014	.907	-0.007
hit2—hit4	1.183	.277	0.066
hit2—hit6	1.396	.237	-0.072
hit3—hit5	5.606	.018	0.144
hit4—hit6	9.762	.002	0.191
hit5—hit7	18.137	.000	0.260

doi:10.1371/journal.pone.0136524.t006

game (complexity and size of action space) as well as individual characteristics of the subjects (gender, CRT score) and response time (Since games from a given subject are likely correlated, we also ran a random effects logistic regression. However, these results did not differ substantially). We find that the complexity of a hit game is the strongest predictor for success in solving it. The effect of the size of the strategy space is a nice indicator of guessing. As already indicated by our previous analysis, response time only matters for problems of complexity 3. Problems of complexity 2 seem to be trivial, i.e., there is no benefit in spending time on thinking about them. Problems of complexity 4 or 5 are overcomplex, i.e., thinking about these problems is just a waste of time. For simplicity, our models only estimate a global effect of non-intuitive thinking (CRT score). We find that non-intuitive thinkers fare better in hit games. However, nonreported analyses show that CRT score matters for simple problems but not for complex problems. As a consequence, the effects of CRT score in [Table 5](#) are quite modest. Note also that intuitive thinkers take less time to make first decisions in hit games (about 10.3 sec. per CRT point, $p = 0.013$; OLS-regression), but this does not exhaust the effects of CRT score. Finally, we observe that males do better in the hit game (i.e. they solve on average 0.38 more hit games, $p = 0.006$; t-test). However, these differences vanish if we take into account that males take more time to think about their first decisions (i.e. on average they take 30 percent more time than females, $p = 0.001$; t-test) and do not rely on intuitions as much as females do (see Sect. 4.1).

4.3 The dirty faces game

Recall that our prime interest regarding the df game is to compare the HU version with the AI version. This allows us to separate the effects of cognitive ability to engage in iterated reasoning from the expectation that the co-players engage in iterated reasoning. More specifically, we want to estimate the fraction of observed behavior in df games which can be explained by the 'theory' that common knowledge of rationality provides. Of course, for any specific dirty faces game we don't need common knowledge of rationality. It suffices that a statement of finite length of the form 'Everybody knows that everybody knows that everybody knows. . . . that everybody is rational.' is true. Common knowledge of rationality is needed to guarantee the solution for any dirty faces game. In the AI version, the only part of the theory that can fail is the subject's rationality, whereas in the HU version common knowledge of rationality might fail additionally because the subjects lack necessary higher order beliefs in the rationality of the co-players. Of course, realistically we expect that beliefs do not fail only on higher levels, but on the most basic level, i.e., we expect that subjects do not believe in the rationality of the co-players. In this sense, the comparison of the rates of behavior which can be explained by common knowledge of rationality (henceforth, ckr-behavior) between the AI and the HU version allows us to estimate the relative importance of ability to engage in iterative reasoning and that of beliefs in the performance of one's co-players.

However, the fact that errors of co-players occur in the HU version but not the AI version requires some attention. The basic question is whether an error of a co-player is observable for a player or not. If an error occurred but is not observable by the player, she can still act rationally on the belief that her co-players are rational, and on subsequent higher order beliefs of rationality. Hence, her behavior can still be reasonably judged against the standards of ckr-behavior. Note that ckr-behavior under conditions of unobservable errors of co-players might involve announcing a type which contradicts her factual type. To see this, consider a two-person df game with one X-player and one O-player. If the X-player erroneously announces U on the first turn, an O-player believing in the rationality of her co-player should announce X on the second turn. In the following, the notion of individually correctly solved df games does not

Table 7. Solved df games by complexity, version, and player's type. Observations in parenthesis.

Complexity	AI version			HU version			
	by type		total	by type		total	
	O	X		O	X		
1		.95 (269)	.95 (269)		.78 (111)	.78 (111)	
2		.57 (256)	.44 (261)	.49 (538)	.83 (314)	.65 (347)	.70 (691)
3		.32 (256)	.11 (262)	.21 (538)	.46 (160)	.14 (390)	.22 (575)
4		.20 (246)	.05 (245)	.12 (538)		.08 (333)	.07 (382)
5						.04 (27)	.03 (34)

doi:10.1371/journal.pone.0136524.t007

refer to factual correctness, but to ckr-behavior in the sense described (Alternatively, we could drop these observations from our data set. However, since unobservable errors occur frequently, we would lose a considerable amount of our data).

Turning to observable errors, two types of errors have to be distinguished, i.e. errors by X-players and errors by O-players. Preliminary analysis of our data showed that cases with observed errors of O-players are very similar to cases without observable errors in terms of the estimated fraction of ckr-behavior. Hence, we include these cases in our analysis. Matters are different for errors by X-players. In most of these cases, theory does not provide a satisfactory solution in terms of ckr-behavior, which is why we exclude them from our analysis.

We now turn to the analysis of observed behavior in df games. Subjects correctly solved 36.6 and 41.9% of the AI and HU games respectively, refuting again the idea of perfectly rational actors. In Table 7 we categorize these results by iteration steps required and subject type (Subjects who announced their type before they could logically deduce it (i.e. guessing), cannot be assigned to a type. However, these subjects are included in the 'total' columns). We observe the same pattern in both versions: The more iteration steps involved, the lower the fraction of subjects that could solve the puzzle. In addition we see that it was more complicated for subjects to solve a game when they were an X-type. Note that this finding runs counter to theoretical predictions. It is plausible that this effect is due to different degrees of salience regarding the informational value of co-players' decisions. If a subject observes that all X-types announce X, this is more thought provoking (Why do they know their types already?) than if the X-types signal U (Why don't they know their types yet?).

At this point we want to stress that in comparison to our discussion of the observed behavior in the hit game, guessing only plays a minor role in the df game. Due to our design, guessing was less profitable than simply picking U each round. Our payoff structure implies that guessing gets even more unattractive on later turns of a round. Hence, looking at early announcements of types which cannot be justified by logical deduction provides a good estimate for the prevalence of guessing in our data on the df game. We find that about 5% of individual plays of df games involve guessing. Note that, in our descriptive statistics announcements of types before these types can actually be deduced logically are classified as non-ckr-behavior.

For convenience, the situations a subject is confronted with will further be labeled mT , where m denotes the number of observable X-types and T denotes the type of the subject. For

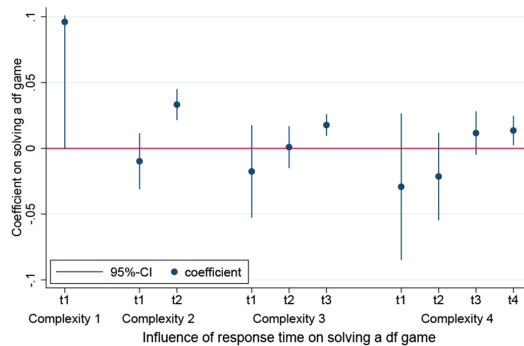


Fig 4. Effect of response time on solving a df game.

doi:10.1371/journal.pone.0136524.g004

simplicity the number of observable O-types is neglected. Interestingly, subjects perform worse in the HU version of the 0X game than in both the AI version of the 0X game as well as in the IO game. This is striking because it is the only situation where beliefs in the rationality of co-players is completely irrelevant. We suggest that this effect is due to some ‘play instinct’ (i.e. ludic drive) to trick their co-players to erroneously announcing X on the second turn. In fact the trick worked really well, since 71 of 92 co-players in these situations actually announced X on the second turn. Ironically, all but 1 of 24 trickers announced X on the second turn as well. In a nutshell, the trickers traded a five percent chance of winning EUR 0.50 for the pleasure of outwitting their co-players ([35], who raffle 75 Australian dollars per person, observe no single case where the 0X problem has not been solved). Also, subjects did not trick the algorithm perhaps because there is no joy in fooling machines.

Similarly to our analysis of the hit game, we now turn to the influence of response times in the df game. Fig 4 plots the coefficients and 95% confidence intervals of response times as estimated by random effects logistic regression models. In these models, the probability of solving a df game of a specific complexity is regressed on the amounts of time subjects invest on each turn and control variables (We deal with the effects of these control variables in a subsequent paragraph). Note that both AI and HU versions of the df game as well as both treatments are pooled in these regressions. Also, note that the theoretical solution based on common knowledge of rationality involves that a player in a game of complexity c (recall, c equals the number of observed X-types + 1) announces U in the first $c - 1$ turns and her type on turn c . Hence, the graph depicts c coefficients and confidence intervals for games of complexity c . Finally, we remark that games of complexity 5 only occurred in the HU version, and in only 34 cases no critical error by a co-player destroyed the solvability of the game for the player (see Table 7). Since these cases do not suffice for regression analyses, games of complexity 5 are not included in our graph.

A straightforward conclusion emerges from the analysis of response times. Response time only matters if invested on the crucial turn, i.e. the turn on which it is de facto possible to logically deduce one’s own type. It is worth noting that subjects invest considerable amounts of time on the noncrucial turns (i.e. on average about 16 seconds), but these investments have no payoff. This finding, in a sense, parallels our findings with respect to the hit game. In the hit

Table 8. Solved df games by treatment. Observations in parenthesis.

Complexity	AH treatment				HA treatment			
	AI version		HU version		HU version		AI version	
	O	X	O	X	O	X	O	X
1		.95 (193)		.77 (79)		.81 (32)		.95 (76)
2	.48 (187)	.40 (188)	.88 (233)	.72 (262)	.73 (91)	.40 (85)	.80 (72)	.52 (73)
3	.32 (184)	.11 (187)	.49 (152)	.17 (256)	.00 (8)	.09 (134)	.35 (72)	.11 (75)
4		.22 (173)		.07 (172)		.12 (223)		.00 (73) (73)
5				.06 (16)		.00 (11)		

doi:10.1371/journal.pone.0136524.t008

game, response time only affected games of complexity 3. Regarding the df game, we find that in situations in which the announcements of the co-players have not yet allowed the deduction of one's own type and hence situations which involve contingent thinking (e.g. 'If all the X-players announce X in turn 3, I will announce O in turn 4, because [...]'), response times have no effect. To us it seems that this means that subjects do not engage or do perform badly in this kind of contingent and hence complex thinking tasks. In contrast to these situations in which only thinking involving contingencies is useful for the solution of the problem, on the crucial turn all the 'facts' regarding the types of the other players are on the table. That is, on the crucial turn no contingent thinking is involved the subjects merely have to properly infer their own types from the available information.

It is time to turn to the main research question of our paper. What factor accounts more for the observed limits in performed iterated reasoning: Limited cognitive abilities to engage in iterated reasoning or lack of beliefs in the abilities of the co-players? Table 8 shows the observed frequencies of ckr-behavior for both AI and HU versions of the df game and both treatment conditions. Most importantly, there is no evidence that the beliefs of the subjects regarding the rationality of their co-players is of any importance in the df game. That is, with some occasional exceptions, the frequency of ckr-behavior is not greater in the AI version than in the HU version of the various df games. In fact, for more experienced subjects, i.e. comparing the HU version in the AH treatment with the AI version in the HA treatment, Table 8 even suggests that in the HU version there is more ckr-behavior than in the AI version. However, this might as well be due to learning effects which might depend on the order of versions. That is, it is plausible that subjects have better chances of learning the game while playing with algorithms than while playing with human co-players. For example, 3O situations did not arise in the HU version of the HA treatment (because of errors by co-players) and hence subjects could not gain any experience for these situations in the HA treatment.

Both the unimportance of beliefs in the rationality of the co-players as well as the treatment effects on learning can also be seen from the *i*-index on the df game (Table 9). Interestingly, the *i*-index is generally higher in the HU version than in the AI version. Hence, beliefs in the rationality of the co-players seem to play no role in the df game. Note that the *i*-index of the AI version in the AH treatments roughly equals the *i*-index of the HU version in the HA treatments.

Table 9. The *i*-index for the df games by treatment and version.

<i>i</i> -index	treatment					
	both		AH		HA	
	AI	HU	AI	HU	HU	AI
0	.05	.09	.05	.09	.08	.05
1	.64	.40	.70	.31	.63	.49
2	.25	.43	.19	.50	.25	.38
3	.06	.05	.05	.05	.04	.08
4	.01	.03	.01	.04	.00	.00

doi:10.1371/journal.pone.0136524.t009

So, for inexperienced subjects it does not seem to be too important whether the co-players are algorithms or humans. However, as already indicated, learning seems to be easier if playing with rational co-players. This can be seen from the fact that the *i*-index of the HU version of the AH treatment puts more weight on higher indices than the *i*-index of the AI version of the HA treatment.

Since we have two *i*-indices for each subject on the df game, i.e. one index for both the AI and the HU version, and these indices refer to chronologically ordered behavior, we can check learning on an individual level. That is, we can look at the proportion of subjects who hold constantly of their level, improve their level, or degenerate. In the HA treatment 58% of subjects hold their level, 32% improve, and 10% actually manage to decrease their level. In the AH treatment, 45% of subjects hold their level, 44% improve their level, and 11% decrease their level. Again, this finding suggests that learning is more efficient in the AH treatment, in which the first part of the treatment confronts the players with rational co-players.

To back up our impressions from descriptive statistics regarding the role of beliefs in the rationality of the co-players and the treatment effects of learning, we estimated a family of random effects logistic regressions. The dependent variable in these regressions is the probability of ckr-behavior in the df game. Table 10 shows three regressions; the first deals with structural variables of the df game, the second adds variables related to treatment and procedure. The final model also incorporates individual characteristics of the subjects, in particular CRT score and hit score.

First of all, structural features of the df game have no surprising effects against the background of our descriptive findings. That is, complexity and being an X-type diminish the probability of ckr-behavior. Notably, the number of co-players does not. This finding already suggests that beliefs about the rationality of the co-players might be empirically irrelevant. The second block of variables teach important lessons. Most notably, it does not matter empirically whether the subjects play with possibly irrational co-players or with rational algorithms. This answers the major motivational question of this paper: Beliefs about the rationality of co-players are irrelevant for the form of iterated reasoning involved in the df game. In addition, we see two kinds of learning effects. On the one hand, experience in the df game benefits ckr-behavior. On the other hand, there is an additional treatment effect, i.e., ckr-behavior is more common when subjects learned the game by playing with algorithms instead of humans first. This interpretation suggests an interaction effect between experience (number of rounds already played) and treatment. Admittedly this interaction effect does not gain significance, which is why it is not reported in these models. Still, we feel that our interpretation is plausible in view of the reported descriptive statistics (see Table 9).

Table 10. Random effects logistic regressions on solving a df game.

<i>df game solved</i>	model			
	(1)	(2)	(3)	(4)
Complexity	-2.023**	-2.319**	-2.319**	-2.318**
Own type (1 = X)	-1.181**	-1.141**	-1.140**	-1.143**
Group size	-0.062	-0.141	-0.127	-0.116
Round		0.130**	0.130**	0.129**
Version (1 = HU)		-0.060	-0.060	-0.058
Treatment (1 = HA)		-0.426*	-0.389*	-0.410*
Response time (sec)		0.019**	0.019**	0.018**
Gender (1 = male)			0.427**	0.210
CRT score				0.368**
Hit score				0.216*
Constant	3.697**	3.093**	2.859**	1.805**
Observations	3480	3480	3480	3480

* $p < 0.05$

** $p < 0.01$

doi:10.1371/journal.pone.0136524.t010

Finally, as already reported, the time subjects invest during the crucial turn benefits ckr-behavior. Model 3 and 4 adds personal characteristics, i.e. gender, CRT and hit score. Similarly to our finding in the hit game, males do better in the df game, although once again these difference can be explained by the fact that males take more time to come to decisions on the crucial turn and faring better in the CRT. More important than the role of gender is the question of whether our experiments provide evidence for some form of generalized cognitive skill in iterated reasoning. And indeed, we find that subjects who performed better in hit games, i.e. iterated reasoning in the form of backward induction, show more affinity to ckr-behavior in the df game. Finally, intuitive thinkers do worse in df games, as they did in hit games.

5 Conclusion

This paper provides experimental evidence on iterated reasoning in games. Three main lessons emerge. First, subjects who do better in backward induction also do better in problems involving interactive knowledge. This suggests that there might exist some generalized ability to engage in iterated reasoning. Second, due to the fact that we took multiple measurements of both forms of iterated reasoning under consideration and also controlled for learning effects, we were able to provide quite reliable measurements of subjects' skill. In comparison to the literature, our estimates of subjects' skill in the hit game [36] and the df game are rather pessimistic [19]. Third and most importantly, our design sheds light on the question of which factor—cognitive ability or beliefs in the abilities of one's co-players—is more important to explain the small amounts of iterated reasoning observed in the literature. We find that in the df game beliefs in the rationality of the co-players are completely irrelevant. In addition to these substantive insights, on methodical grounds this paper exemplifies the usefulness of response time analysis to validate estimates of subjects' abilities in cognitively demanding tasks [25, 26].

Clearly, our main finding regarding the relative importance of cognitive abilities and beliefs in the rationality of co-players cannot be easily generalized to other types of games and subjects. It might well be that the df game is far too complex to allow more or less inexperienced

subjects to engage in reasoning about the rationality of their co-players. It is therefore important to use similar designs with more experienced subjects or simpler games [37]. In any case, we feel that more experimental efforts should be directed at assessing whether and to what extent humans actually take into account the perceived weaknesses in the rationality of their co-players, because this question seems vital for an empirically oriented game theory, i.e. a theory of interactive decision making that actually captures how real actors play games.

Supporting Information

S1 Instructions. Instruction materials given to our subjects. Translated from German. (PDF)

S1 Appendix. Extended descriptions of the hit game and the dirty faces game. (PDF)

S1 Dataset. Used data in csv format. (CSV)

Author Contributions

Conceived and designed the experiments: SG AT. Performed the experiments: SG AT. Analyzed the data: SG AT. Contributed reagents/materials/analysis tools: SG AT. Wrote the paper: SG AT.

References

1. Devetag G, Warglien M. Games and Phone Numbers: Do Short-Term Memory Bounds Affect Strategic Behavior? *The Economic Psychology of Herbert A. Simon*. *Journal of Economic Psychology*. 2003; 24(2):189–202. Available from: <http://www.sciencedirect.com/science/article/pii/S0167487002002027>
2. Gneezy U, Rustichini A, Vostroknutov A. Experience and insight in the Race game. *Journal of Economic Behavior and Organization*. 2010; 75(2):144–155. doi: [10.1016/j.jebo.2010.04.005](https://doi.org/10.1016/j.jebo.2010.04.005)
3. Blume A, Gneezy U. Cognitive forward induction and coordination without common knowledge: An experimental study. *Games and Economic Behavior*. 2010; 68(2):488–511. doi: [10.1016/j.geb.2009.07.011](https://doi.org/10.1016/j.geb.2009.07.011)
4. Rubinstein A. *Modeling Bounded Rationality*. Cambridge and London: MIT Press; 1998.
5. Rubinstein A, Osborne MJ. Games with Procedurally Rational Players. *American Economic Review*. 1998; 88(4):834–847.
6. Spiegel R. Equilibrium in Justifiable Strategies: A Model of Reason-Based Choices in Extensive-Form Games. *Review of Economic Studies*. 2002; 69(3):691–706. doi: [10.1111/1467-937X.00222](https://doi.org/10.1111/1467-937X.00222)
7. Goeree JK, Holt CA. Ten Little Treasures of Game Theory and Ten Intuitive Contradictions. *American Economic Review*. 2001; 91(5):1402–1422. doi: [10.1257/aer.91.5.1402](https://doi.org/10.1257/aer.91.5.1402)
8. Grosskopf B, Nagel R. The two-person beauty contest. *Games and Economic Behavior*. 2008; 62(1):93–99. doi: [10.1016/j.geb.2007.03.004](https://doi.org/10.1016/j.geb.2007.03.004)
9. Huck S, Weizsäcker G. Do players correctly estimate what others do?: Evidence of conservatism in beliefs. *Journal of Economic Behavior and Organization*. 2002; 47(1):71–85. Available from: <http://www.sciencedirect.com/science/article/pii/S0167268101001706> doi: [10.1016/S0167-2681\(01\)00170-6](https://doi.org/10.1016/S0167-2681(01)00170-6)
10. Aumann R, Brandenburger A. Epistemic Conditions for Nash Equilibrium. *Econometrica*. 1995; 63(5):1161. doi: [10.2307/2171725](https://doi.org/10.2307/2171725)
11. Stahl DO, Wilson PW. Experimental Evidence on Players' Models of Other Players. *Journal of Economic Behavior and Organization*. 1994; 25(3):309–327. doi: [10.1016/0167-2681\(94\)90103-1](https://doi.org/10.1016/0167-2681(94)90103-1)
12. Stahl DO, Wilson PW. On Players' Models of Other Players: Theory and Experimental Evidence. *Games and Economic Behavior*. 1995; 10(1):218–254. doi: [10.1006/game.1995.1031](https://doi.org/10.1006/game.1995.1031)
13. Nagel R. Unraveling in Guessing Games: An Experimental Study. *American Economic Review*. 1995; 85(5):1313–1326.

14. Camerer CF, Ho TH, Chong JK. A Cognitive Hierarchy Model of Games. *The Quarterly Journal of Economics*. 2004; 119(3):861–898. doi: [10.1162/0033553041502225](https://doi.org/10.1162/0033553041502225)
15. Bosch-Domènech A, Montalvo JG, Nagel R, Satorra A. A finite mixture analysis of beauty-contest data using generalized beta distributions. *Experimental Economics*. 2010; 13(4):461–475. doi: [10.1007/s10683-010-9251-7](https://doi.org/10.1007/s10683-010-9251-7)
16. Penczynski SP. Strategic Thinking: The Influence of the Game; 2014.
17. Arad A, Rubinstein A. Multi-dimensional iterative reasoning in action: The case of the Colonel Blotto game. *Journal of Economic Behavior and Organization*. 2012; 84(2):571–585. doi: [10.1016/j.jebo.2012.09.004](https://doi.org/10.1016/j.jebo.2012.09.004)
18. Arad A, Rubinstein A. The 11–20 Money Request Game: A Level- k Reasoning Study. *American Economic Review*. 2012; 102(7):3561–3573. doi: [10.1257/aer.102.7.3561](https://doi.org/10.1257/aer.102.7.3561)
19. Weber RA. Behavior and Learning in the “Dirty Faces” Game. *Experimental Economics*. 2001; 4(3):229–242. doi: [10.1023/A:1013217320474](https://doi.org/10.1023/A:1013217320474)
20. Dufwenberg M, Sundaram R, Butler DJ. Epiphany in the Game of 21. *Journal of Economic Behavior and Organization*. 2010; 75(2):132–143. doi: [10.1016/j.jebo.2010.03.025](https://doi.org/10.1016/j.jebo.2010.03.025)
21. Agranov M, Caplin A, Tergiman C. Naïve play and the process of choice in guessing games. Working Paper; 2012.
22. Burnham TC, Cesarini D, Johannesson M, Lichtenstein P, Wallace B. Higher cognitive ability is associated with lower entries in a p -beauty contest. *Journal of Economic Behavior and Organization*. 2009; 72(1):171–175. doi: [10.1016/j.jebo.2009.05.015](https://doi.org/10.1016/j.jebo.2009.05.015)
23. Carpenter J, Graham M, Wolf J. Cognitive ability and strategic sophistication. *Games and Economic Behavior*. 2013; 80(0):115–130. Available from: <http://www.sciencedirect.com/science/article/pii/S0899825613000365> doi: [10.1016/j.geb.2013.02.012](https://doi.org/10.1016/j.geb.2013.02.012)
24. Agranov M, Potamites E, Schotter A, Tergiman C. Beliefs and endogenous cognitive levels: An experimental study. *Games and Economic Behavior*. 2012; 75(2):449–463. doi: [10.1016/j.geb.2012.02.002](https://doi.org/10.1016/j.geb.2012.02.002)
25. Rubinstein A. Instinctive and Cognitive Reasoning: a Study of Response Times. *The Economic Journal*. 2007; 117(523):1243–1259. doi: [10.1111/j.1468-0297.2007.02081.x](https://doi.org/10.1111/j.1468-0297.2007.02081.x)
26. Rubinstein A. Response time and decision making: An experimental study. *Judgment and Decision Making*. 2013; 8(5):540–551.
27. Bayer RC, Renou L. Logical Omniscience at the Laboratory. Working Paper; 2009.
28. Frederick S. Cognitive Reflection and Decision Making. *The Journal of Economic Perspectives*. 2005; 19(4):25–42. Available from: <http://www.jstor.org/stable/4134953> doi: [10.1257/089533005775196732](https://doi.org/10.1257/089533005775196732)
29. Kahneman D. *Thinking, fast and slow*. London: Penguin; 2012.
30. Stanovich KE, West RF. Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*. 2000; 23(5):645–665. doi: [10.1017/S0140525X00003435](https://doi.org/10.1017/S0140525X00003435) PMID: [11301544](https://pubmed.ncbi.nlm.nih.gov/11301544/)
31. Osborne MJ, Rubinstein A. *A Course in Game Theory*. Cambridge: MIT Press; 1994.
32. Bayer RC, Chan M. The Dirty Faces Game Revisited. Working Paper; 2007.
33. Bock O, Nicklisch A, Baetge I. hroot: Hamburg recruitment and organization online tool; 2012.
34. Fischbacher U. z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*. 2007; 10(2):171–178. doi: [10.1007/s10683-006-9159-4](https://doi.org/10.1007/s10683-006-9159-4)
35. Bayer RC, Renou L. Facing Man or Machine When Do Humans Reason Better? Working Paper; 2013.
36. Levitt SD, List JA, Sadoff SE. Checkmate: Exploring Backward Induction among Chess Players. *American Economic Review*. 2011; 101(2):975–990. doi: [10.1257/aer.101.2.975](https://doi.org/10.1257/aer.101.2.975)
37. Palacios-Huerta I, Volij O. Field Centipedes. *American Economic Review*. 2009; 99(4):1619–1635. doi: [10.1257/aer.99.4.1619](https://doi.org/10.1257/aer.99.4.1619)

A Note on Disbelief in Others regarding Backward Induction

Andreas Tutić, Sascha Grehl
Games 8(3), 33 (2017).

Communication

A Note on Disbelief in Others regarding Backward Induction

Andreas Tutić^{1,2,*} and Sascha Grehl¹

¹ Institute of Sociology, Leipzig University, 04107 Leipzig, Germany; sascha.grehl@uni-leipzig.de

² Institute of Sociology, University of Bern, 3012 Bern, Switzerland

* Correspondence: andreas.tutic@sozio.uni-leipzig.de

Received: 17 July 2017; Accepted: 4 August 2017; Published: 8 August 2017

Abstract: We present experimental results on the role of beliefs in the cognitive ability of others in a problem involving backward induction. Using a modified version of the so-called race game, our design allows the effects of a player's own inability to perform backward induction to be separated from the effects of her disbelief in the ability of others to do so. We find that behavior is responsive to the dependence on others who might fail in backward induction as well as information regarding their backward induction skills.

Keywords: backward induction; iterative thinking; beliefs

1. Introduction

By now, extensive experimental research has driven home the point that the behavior of inexperienced players in novel strategic situations is in stark contrast to game-theoretical predictions [1–11]. Several factors can be identified as potential causes for out-of-equilibrium behavior in “initial responses” [12]. These include unobservable other-regarding preferences as well as a general inability to understand the rules of the game [13,14]. Special attention has been directed at studying two factors of out-of-equilibrium behavior; i.e., bounded rationality and disbelief in the cognitive abilities of co-players. Aumann, for instance, has drawn upon arguments from epistemic game theory to show that typical observations regarding the centipede game can be partially explained by a “slight” amount of irrationality and a relaxation of common knowledge of rationality [15,16]. The underlying intuition is clear: Playing according to equilibrium predictions might be unwise when paired with irrational players or players who believe in the irrationality of their co-players, or mutual knowledge of rationality of higher order is violated.

A considerable amount of research has been conducted to assess the relative importance of disbelief in the cognitive abilities of co-players (for brevity: disbelief in others) in explaining out-of-equilibrium behavior. Unfortunately, the resulting experimental evidence is mixed. That is, some studies find evidence for disbelief in others in the centipede game [9] or the guessing game [11], other studies find no effects [17,18]. Against this background, this paper sheds light on the question of whether and to what extent subjects take the cognitive skills of their co-players into account when dealing with problems involving backward induction. We use an extended version of the so-called race game [19–21] (i.e., a two-person zero-sum game in which one player can enforce a win by playing a weakly dominant strategy, which can be identified by applying backward induction). The basic race game is extended by a stage in which the players choose among two payoff options, which allows measurement of their confidence in winning the respective race game. Treatments differ regarding whether subjects play solo or are matched in teams of two when playing against an algorithm which was programmed to mimic a perfectly rational co-player with the sole motivation of winning the game. Treatments also differ with respect to the available information regarding the skills in backward

induction of the respective team member. We find clear evidence for disbelief in others: that is, subjects condition their behavior on the information they have regarding the skills in backward induction of their team member. We also observe that subjects generally overestimate the strategic skills of their team member; i.e., subjects having no information regarding their co-player behave identically to subjects who know that they are teamed up with a (relatively) skilled team member.

2. Experimental Procedure and Design

The experiment was conducted in German at the experimental laboratory of the Leipzig University, Germany, in spring 2015.¹ A total of 188 subjects participated in 15 sessions, which lasted for about 60 min. The sample consisted of students and seven former students, who had graduated only recently. The mean age was 24, and females were slightly overrepresented (66%). On average, subjects earned 9.74 Euro.

Subjects played a so-called extended race games (hereafter: ERG). The basic race game is an extensive two-person game in which one player can enforce a win by playing a weakly dominant strategy that can be identified by applying backward induction [19–21]. At the beginning of a race game, a certain number of balls are laid out. The first player then has to choose a number of balls that will be removed. Right afterward, the second player gets to choose a number of balls to be removed. The two players alternate in this manner until a player removes the last ball and by doing so wins the game. A specific basic race game is characterized by a triple (m, ℓ, u) , in which m denotes the number of initial balls and $0 < \ell < u$ describe the lower and upper bounds of balls that can be removed on each turn.

Similar to the literature [21], in each game subjects knowingly play against an algorithmic co-player (AI) which was programmed to mimic a perfectly rational co-player with the sole motivation of winning the game. In each game, subjects start in a position in which a win was enforceable by playing a backward-induction strategy.²

The extension of the basic race game comprises two dimensions. First, we measure subjects' confidence to win each specific game. To achieve this, subjects are informed about the game parameters before the game starts and have to choose among two payoff options which differ with respect to the chances of obtaining a fixed monetary price of 80 cents in the case of a win or a loss. The payoff options are presented to the subjects as option A and B, respectively. Option A results in a 100% chance of obtaining the fixed price if the current game is won, but results in a 0% chance of obtaining the price if the game is lost. Option B on the other hand, offers a 70% chance in case of a win and a 30% chance in case of a loss. Clearly, option A is more attractive if the subject believes that winning is more likely than losing, whereas option B is more attractive in case the subject believes that losing is more likely than winning. Note that subjects had two minutes for choosing a payoff option and winning the game, otherwise they lost automatically and got no monetary reward for this game.

Second, we assess the impact of disbelief in others on subjects' behavior by varying the winning condition. Specifically, subjects play two series of seven distinct ERGs (The games implemented can be seen in Table 1).³ In the first series (ERG1), subjects win a game if they take the last ball in the play against the AI. The idea behind this series is simply to measure subjects' skills in backward induction. We will henceforth refer to the number of games won in this series as the subject's BI-score. In the second series (ERG2), subjects are randomly allocated into one of three treatments (see Table 2). In the single treatment, which serves as the control treatment, the winning condition is identical to the first series. In the team treatment, subjects are matched in teams of two; this matching was done randomly

¹ Subjects were recruited via the internet recruitment tool hroot [22] and the whole experiment was conducted with the software z-Tree [23].

² The instructions can be found in the Appendix.

³ Due to our research interest in initial responses, we aim at impeding learning as much as possible; therefore, each game is unique.

and anew for each game in the second series. Neither team member directly interacts with the other. However, in this treatment, each subject wins if and only if both the subject and her team member beat the AI in their respective game. Subjects were made aware that both team members play the same game against the AI. Importantly, in this treatment, the beliefs of the subjects regarding the cognitive skills of their team members are vital for their choice between the payoff options. Note also that only the first-order belief in the rationality of the team member is under scrutiny; i.e., it does not matter for ego’s choice of payoff options whether the team member believes in ego’s rationality or not (nor does any higher-order belief matter). The team-info treatment is similar to the team treatment, but differs with respect to the available information regarding the skills in backward induction of the team members; i.e., in the team treatment, subjects are only informed that they are matched in teams, whereas in the team-info treatment, subjects can obtain information concerning the skills of their team member in backward induction (precisely, the team member’s BI-score). In order to observe whether subjects in the team-info treatment care about this information, the info is initially hidden until subjects manually reveal it. To suppress learning, subjects in the team treatment as well as in the team-info treatment did not receive any feedback regarding the performance of their team members (except their respective BI-score) until the second series was finished.⁴

Table 1. The two extended race game (ERG) series implemented.

Series	Parameter	Game Number								
		0*	1	2	3	4	5	6	7	
FirstERG1	ℓ	1	2	1	2	1	2	1	2	
	u	-----			3	-----				
	m	6	6	11	11	13	13	18	18	
SecondERG2	ℓ	1	2	1	2	1	2	1	2	
	u	-----			4	-----				
	m	7	10	12	13	14	16	17	19	

Notes: * practice round (not included in the analysis).

Table 2. Experimental design and number of subjects (number of observations in parentheses).

Treatment	First Series	Second Series	Subjects
Single	ERG1 (308)	ERG2 single (308)	N = 44
Team	ERG1 (448)	ERG2 team (448)	N = 64
Team-info	ERG1 (560)	ERG2 team-info (560)	N = 80

The gist of this design is as follows: if subjects are influenced by disbelief in others, we should observe the choice of option B more frequently in the team and team-info treatment than in the single treatment. Further, subjects in the team-info treatment should choose option A more often, the higher the BI-score of their team member.

3. Results

Concerning subjects’ abilities in backward induction, we observe that in both series, less than 20% of the subjects won more than half of the games. Additionally, the number of games won in the first series (i.e., the BI-Score) is on average 2.43, backing the finding that applying backward induction is troublesome in initial response [20].

To answer the question of whether disbelief in others affects decisions in backward induction problems, we look at treatment effects regarding the choice of payoff options in the second series.

⁴ Of course, subjects could always observe their own performance in the extended race games.

Figure 1a shows the proportion of choices of option B (henceforth: B-choices) in each game of the second series, separated by treatment. We observe the following strict monotonic order: in each particular game, option B was chosen most frequently in the team-info treatment, followed by the team treatment, and least frequently in the single treatment. In addition, we observe that the differences between the treatments get smaller in later (and, by design, more complex) games. This makes sense, since the more complex the game, the less confident subjects should be in their capacity to win their own game against the AI, and hence the less important are their doubts in the abilities of their team member. When pooled, these differences are highly significant between the team-info treatment and the team as well as the single treatment (both: $p < 0.002$; χ^2 -test), and weakly significant between the team and the single treatment ($p < 0.065$; χ^2 -test). Hence, we conclude that disbelief in others influences the subjects.

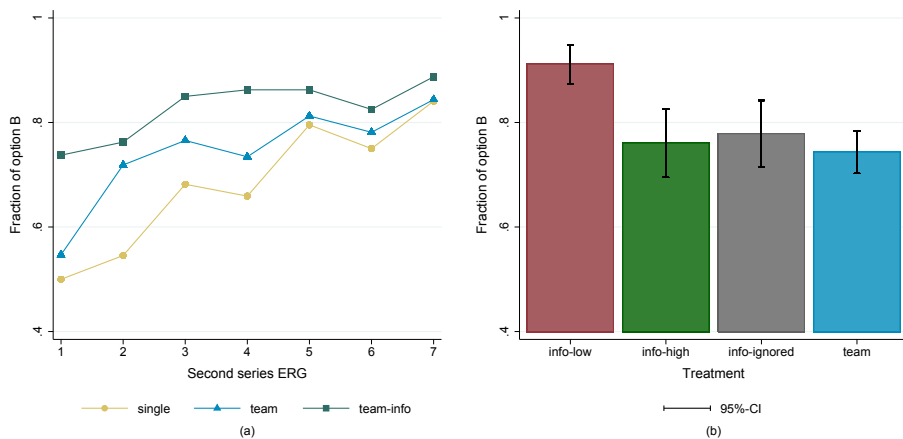


Figure 1. The proportion of B-choices in the 2nd series (a) for each game by treatment and (b) pooled by information (with 95%-C.I.).

Regarding the effect of information on the co-player’s backward induction skill, panel (b) of Figure 1 depicts the proportion of B-choices in the team treatment as well as in the team-info treatment. With respect to the team-info treatment, the figure differentiates between three kinds of subjects: those subjects who decide to ignore the information about the BI-score of their team member (info-ignored), as well as two groups of subjects who examine this information and whose team member belongs either to the 50% of top performers (info-high) or 50% of low performers (info-low). Among those subjects who have a look at the BI-score of their team member, those who are paired with a relatively low-skilled co-player chose option B significantly more often than those paired with a high-skilled co-player ($p < 0.001$; χ^2 -test). Further, we observe that subjects receiving the information of a relatively high BI-score do not show behavior that differs significantly from that of subjects who did not have any information either by treatment condition or by simply ignoring the information (info-high vs. team: $p < 0.664$, info-high vs. info-ignored: $p < 0.698$; χ^2 -test). This finding suggests that subjects tend to overestimate their team members’ abilities, which explains why disbelief in others shows more of an impact in the team-info than in the team treatment.

Finally, we estimate three random effects logit regressions. In each model, the dependent variable is a dummy indicating whether option B is chosen. The first model (see Table 3) includes basic treatment conditions as well as a dummy indicating whether the game at hand stems from the first or from the second series. We observe that playing in teams, as well as the availability of information regarding the team member’s performance, significantly increase the probability of choosing option B. In addition, we observe that subjects tend to choose option B more frequently in the second series

than in the first series. This is quite reasonable, since games of the second series are cognitively more demanding.

Table 3. Random effects logit regressions of B-choices in both race series.

Choosing Option B?	Model		
	(1)	(2)	(3)
Team condition? (no = 0, yes = 1)	0.579 (0.253) **	0.580 (0.253) **	0.596 (0.274) **
Info condition? (no = 0, yes = 1)	0.768 (0.241) ***	0.468 (0.332)	0.393 (0.345)
Second series? (no = 0, yes = 1)	0.917 (0.193) ***	0.918 (0.193) ***	0.953 (0.206) ***
Info. examined? (no = 0, yes = 1)		2.564 (0.583) ***	2.681 (0.607) ***
If info. examined: Team member's BI-score		−0.812 (0.174) ***	−0.828 (0.181) ***
Male (no = 0, yes = 1)			−0.449 (0.296)
Age (in years)			−0.058 (0.044)
Father's education (1 [low]–6 [high])			0.105 (0.073)
Constant	0.007 (0.135)	0.006 (0.136)	1.126 (1.116)
Observations	2632	2632	2352

Notes: Logit coefficients, standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

In the second model, we add two variables which give a more nuanced picture of the influence of the information regarding the team member's performance. The first one is a dummy which captures whether subjects actually examine the information of their team members' BI-scores (regardless of whether subjects ignore the information or are simply not in an information condition). The second variable is an interaction term that is the product of the aforementioned dummy and the team member's BI-score. This results in a variable ranging from 0 to 5. Most importantly, we observe that subjects who examine the information are more likely to choose option B; however, this effect is mediated by the actual information observed (i.e., the higher the observed BI-score of the team member, the less likely it is that a subject will opt for option B). We find that the odds of choosing option B for subjects who could not or did not look at the information are roughly the same as for subjects who examine the information and observe that their team member has a BI-score of 3. While the effect of the team condition is almost identical in model 1 and 2, the effect of the information condition vanishes to insignificance. This indicates that the behavioral changes between the team and team-info treatments are caused solely by the information.

The third model validates that our findings are robust when controlling for essential demographic characteristics.⁵

4. Conclusions

In this paper, we present experimental evidence on backward induction and shed light on the question of whether and to what extent disbelief in others influences behavior in such a setting. We find clear evidence that disbelief in others affects behavior in problems involving backward induction. In addition, this paper documents that subjects condition their choice on their information regarding the co-player. Interestingly, subjects who have no information about their co-players tend to overestimate their cognitive skills, resulting in a behavior that is similar to subjects who know that they are playing with a relatively highly-skilled co-player.

⁵ The variable "father's education" is ordinal and takes the following values depending on the highest degree of education the subject's father obtained: 1 = Certificate of Secondary Education (Hauptschulabschluss), 2 = General Certificate of Secondary Education (Realschulabschluss), 3 = Restricted qualification for university entrance (Fachschulabitur), 4 = General qualification for university entrance (Abitur), 5 = Bachelor degree, 6 = Master degree (Diplom/Magister). Results do not change if we work with dummies for each type of educational level.

Acknowledgments: Research funding by the German Research Foundation (TU 409/1-1) is gratefully acknowledged.

Author Contributions: Andreas Tutić and Sascha Grehl conceived and designed the experiments; Sascha Grehl performed the experiments; Andreas Tutić and Sascha Grehl analyzed the data and wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Here we provide the instructions regarding the race games (translated from German). Note that the bold text was optional, depending on treatment.

Race Game-Instructions

In this part, you will play against the computer. At the beginning of each round, a certain number of balls are placed on the screen. You and the computer alternately act in turns. In each turn, some balls must be removed. The player who removes the last balls wins the game, the other player loses. The computer is programmed in such a way that it wants to win the game and is therefore planning several steps ahead.

For the removal of the balls, the following rules apply:

- On each turn, no more than 4 balls are allowed to be removed.
- On each turn, at least 1 or 2 balls have to be removed (this can vary depending on the round).

Depending on whether you have removed the last ball or not, you will get a lottery ticket of different quality. **In addition, at the beginning of each round, you form with a randomly selected participant a team. Both you and your team member play separate games against the computer (however, both games will be the same). Only if you and your team member succeed to remove the last ball in both games, you will get the better lottery ticket.**

Before each round you must decide among these two payoff options:

- (A) **You and your team member** take the last ball: You get a lottery ticket that wins with a probability of 100%. Otherwise, you will get a lottery ticket that wins with a probability of 0%.
- (B) **You and your team member** take the last ball: You get a lottery ticket that wins with a probability of 70%. Otherwise, you will get a lottery ticket that wins with a probability of 30%.

A lottery ticket that wins is worth 80 points [authors note: 0.8 Euro]. All lottery tickets will be drawn at the end of this study. **Please note that your choice regarding the payoff options does not affect the payoffs of your team member, and vice versa.**

While you decide for an option, you will also receive information regarding the number of rounds your team member has won the game against the computer in the previous part of this study. To view this information, you must click on the red box in the lower left part of the screen.

In each round, you have 120 s to choose a payoff option and to win the game. When this time is over, you will lose automatically and receive a payoff of 0 points [authors note: 0 Euro] for this round.

First, a practice round will be played, which does not affect your payoffs. Use this round to get an overview. After the practice round another 7 rounds are played, which differ regarding the number of balls at the beginning and the minimal number of balls that have to be removed in a turn. Note that you are always the starting player.

Finally, please note that you also win the game when you have to remove more balls than are currently laid out. This can happen if you have to remove at least 2 balls, but only one ball is present.

Good Luck!

References

- Rosenthal, R.W. Games of perfect information, predatory pricing and the chain-store paradox. *J. Econ. Theory* **1981**, *25*, 92–100.
- McKelvey, R.D.; Palfrey, T.R. An experimental study of the centipede game. *Econometrica* **1992**, *60*, 803–836.
- Nagel, R. Unraveling in guessing games: An experimental study. *Am. Econ. Rev.* **1995**, *85*, 1313–1326.
- Fey, M.; McKelvey, R.D.; Palfrey, T.R. An experimental study of Constant-Sum centipede games. *Int. J. Game Theory* **1996**, *25*, 269–287.
- Ho, T.H.; Camerer, C.F.; Weigelt, K. Iterated dominance and iterated best response in experimental “p-beauty contests”. *Am. Econ. Rev.* **1998**, *88*, 947–969.
- Nagel, R.; Tang, F.F. Experimental results on the centipede game in normal form: An investigation on learning. *J. Math. Psychol.* **1998**, *42*, 356–384.
- Bosch-Domènech, A.; Montalvo, J.G.; Nagel, R.; Satorra, A. One, two, (three), infinity, ... newspaper and lab beauty-contest experiments. *Am. Econ. Rev.* **2002**, *92*, 1687–1701.
- Burnham, T.C.; Cesarini, D.; Johannesson, M.; Lichtenstein, P.; Wallace, B. Higher cognitive ability is associated with lower entries in a p-beauty contest. *J. Econ. Behav. Organ.* **2009**, *72*, 171–175.
- Palacios-Huerta, I.; Volij, O. Field centipedes. *Am. Econ. Rev.* **2009**, *99*, 1619–1635.
- Levitt, S.D.; List, J.A.; Sadoff, S.E. Checkmate: Exploring backward induction among chess players. *Am. Econ. Rev.* **2011**, *101*, 975–990.
- Agranov, M.; Caplin, A.; Tergiman, C. Naive play and the process of choice in guessing games. *J. Econ. Sci. Assoc.* **2015**, *1*, 146–157.
- Crawford, V.P.; Costa-Gomes, M.A.; Iriberri, N. Structural models of nonequilibrium strategic thinking: Theory, evidence, and applications. *J. Econ. Lit.* **2013**, *51*, 5–62.
- Chou, E.; McConnell, M.; Nagel, R.; Plott, C.R. The control of game form recognition in experiments: Understanding dominant strategy failures in a simple two person “Guessing” game. *Exp. Econ.* **2009**, *12*, 159–179.
- Devetag, G.; Warglien, M. Playing the wrong game: An experimental analysis of relational complexity and strategic misrepresentation. *Games Econ. Behav.* **2008**, *62*, 364–382.
- Aumann, R.J. Irrationality in game theory. In *Economic Analysis of Markets and Games*; Dasgupta, P., Gale, D., Hart, O., Maskin, E., Eds.; MIT Press: Cambridge, UK, 1992; pp. 214–227.
- Reny, P.J. Common knowledge and games with perfect information. In *Proceedings of the Biennial Meeting of the Philosophy of Science Association*; The University of Chicago Press: Chicago, IL, USA, 1988; Volume 2, pp. 363–369.
- Georganas, S.; Healy, P.J.; Weber, R.A. On the persistence of strategic sophistication. *J. Econ. Theory* **2015**, *159*, 369–400.
- Levitt, S.D.; List, J.A.; Reiley, D.H. What happens in the field stays in the field: Exploring whether professionals play minimax in laboratory experiments. *Econometrica* **2010**, *78*, 1413–1434.
- Gneezy, U.; Rustichini, A.; Vostroknutov, A. Experience and insight in the race game. *J. Econ. Behav. Organ.* **2010**, *75*, 144–155.
- Dufwenberg, M.; Sundaram, R.; Butler, D.J. Epiphany in the game of 21. *J. Econ. Behav. Organ.* **2010**, *75*, 132–143.
- Brosig-Koch, J.; Heinrich, T.; Helbach, C. Exploring the capability to reason backwards: An experimental study with children, adolescents, and young adults. *Eur. Econ. Rev.* **2015**, *74*, 286–302.
- Bock, O.; Nicklisch, A.; Baetge, I. hroot: Hamburg Recruitment and Organization Online Tool. *Eur. Econ. Rev.* **2014**, *71*, 117–120.
- Fischbacher, U. z-Tree: Zurich Toolbox for Ready-made Economic Experiments. *Exp. Econ.* **2007**, *10*, 171–178.



Status Characteristics and the Provision of Public Goods – Experimental Evidence

Andreas Tutić, Sascha Grehl
Sociological Science 5, 1-20 (2018).

Status Characteristics and the Provision of Public Goods: Experimental Evidence

Andreas Tutić, Sascha Grehl

Leipzig University

Abstract: We present experimental evidence on the effects of status characteristics in problems involving the provision of public goods. According to Status Characteristics Theory (SCT), status differentials affect performance expectations, which in turn affect the power and prestige order in group tasks. Applied to problems of collective action, SCT suggests several intriguing hypotheses (cf. Simpson, Willer, and Ridgeway 2012). Most importantly, the theory proposes that high-status actors show a greater initiative in and also overall contribute more to the provision of public goods than low-status actors. We put this theoretical claim to a strict experimental test, in addition to other hypotheses and conjectures. In our experimental setup, the volunteer's timing dilemma is used as the group task. Three experimental conditions are implemented, which differ with respect to the way status groups are formed on basis of the type of status characteristic. Our results validate the central hypothesis cited above and also lend support to a conjecture regarding the beneficial effects of heterogeneity in status.

Keywords: volunteer's timing dilemma; social status; public goods; small-group research

THE question of if and to what extent status groups differ in prosocial behavior has a long tradition in the sociological literature (e.g., Mauss [1925] 1954; Homans 1974; Coleman 1990). Because recent advances in experimental social science provide more nuanced tools to differentiate and measure diverse forms of prosocial behavior such as altruistic giving or reciprocating fairness, the status–prosociality nexus has received renewed attention (e.g., Simpson and Willer 2015). Using a wide variety of measures of social status and focusing on plenty of theoretical mechanisms linking social status to prosocial behavior, studies find for instance that status matters regarding donations in the dictator game (e.g., Liebe and Tutić 2010; Piff et al. 2010) as well as the placement of trust in the trust game (e.g., Piff et al. 2010). However, the impact of status hierarchies on one particular core dimension of prosocial behavior—that is, collective action and the voluntary provision of public goods (Olson 1965; Ostrom 1990; Heckathorn 1996)—has so far eschewed extensive experimental inquiry.

It has been just recently that Simpson, Willer, and Ridgeway (2012) elucidated how Status Characteristics Theory (SCT) offers a suitable account for theorizing about the effects of status differentials on collective action. SCT posits that diffuse status characteristics such as gender or ethnicity as well as specific characteristics such as reading ability or mathematical intelligence affect behavior in group tasks via the formation of performance expectations (cf. Berger et al. 1977; Berger, Rosenholtz, and Zelditch 1980; Berger, Wagner, and Zelditch 1985). Simpson et al. (2012) argue that SCT applies to problems of collective action and derive the following three hypotheses: (1) high-status actors take more initiative in contributing towards

Citation: Tutić, Andreas, and Sascha Grehl. 2018. "Status Characteristics and the Provision of Public Goods: Experimental Evidence" *Sociological Science* 5: 1-20.


Received: October 30, 2017

Accepted: November 24, 2017

Published: January 4, 2018

Editor(s): Jesper Sørensen, Gabriel Rossman

DOI: 10.15195/v5.a1

Copyright: © 2018 The Author(s). This open-access article has been published under a Creative Commons Attribution License, which allows unrestricted use, distribution and reproduction, in any form, as long as the original author and source have been credited. 

the provision of public goods than low-status actors, (2) high-status actors contribute more towards the provision of public goods than low-status actors, and (3) low-status actors are more eager to match the contributions of high-status actors than vice versa. In addition, Simpson et al. (2012) conjecture that heterogeneity in terms of status characteristics might benefit groups facing problems of collective action since it helps to overcome start-up and free-riding problems.

In this article, we provide experimental evidence on predictions derived from SCT regarding the provision of collective goods. Specifically, we put hypotheses (1) and (2) as well as the conjecture by Simpson et al. (2012) to the test. In our experimental setup, subjects are teamed up in groups of four and confronted with a modified version of the volunteer's timing dilemma (cf. Weesie 1993; Otsubo and Rapoport 2008), in which at least two players have to bear a private cost in order to secure the provision of a public good. Each group is composed of two high-status (so-called stars) and two low-status actors (so-called nonstars). There are three treatments that differ in the way of how subjects are allocated to status groups (i.e., the treatments differ regarding the source of status). In the random treatment, subjects are randomly assigned to the star or nonstar group and, crucially, are informed about the randomness of assignments. In the diffuse treatment, subjects are allocated to status groups based on their subjective social status as measured by a modification of the MacArthur Scale (cf. Adler et al. 2000). In the specific treatment, subjects are told that they have been assigned to status groups based on their performance in a quiz regarding basic understanding of game-theoretic concepts. Yet in fact, they are randomly assigned to status groups.

We find no behavioral differences among status groups in the random treatment. Yet once status groups are composed or allegedly composed based on differences in diffuse or specific status characteristics, high-status actors do show more initiative and do overall contribute more to the provision of public goods than low-status actors. These findings are very much in line with SCT. Also, as conjectured by Simpson et al. (2012), heterogeneity in status characteristics does benefit groups in our experiment in terms of rates of group success in the provision of public goods as well as experimental earnings.

This article extends our knowledge on collective action problems and SCT in three important ways. First, although there is considerable experimental evidence for the validity of core principles of SCT, the larger portion of these findings has been derived from experiments that followed a standardized protocol. This experimental setting involves in a sense fictitious specific characteristics such as contrast sensitivity, meaning-insight ability, relational ability, et cetera as well as specific tasks such as judging the relative size of geometrical figures (Berger et al. 1977: 43–48). Clearly, providing evidence on SCT in other and—from a social-theoretical point of view—more relevant areas of interactive decision-making such as problems of collective action is valuable per se. Second, this is the only study that allows comparing the behavioral effects of different sources of status within a fixed experimental protocol in the context of the voluntary provision of public goods. That is, hitherto scholars have studied the effects of either diffuse status characteristics such as gender (Sell 1997) and scholastic experience (Simpson et al. 2012) or specific characteristics such as performance in general knowledge quizzes (Kumru and

Vesterlund 2010). The question of whether different sources of status affect cooperative behavior in collective action problems in a similar way has consequently eschewed experimental investigation. Hence, our study provides a more reliable basis regarding the extent to which results from prior studies working with diffuse characteristics generalize to specific characteristics and vice versa. Third, our study extends experimental evidence regarding the effects of status characteristics in the voluntary provision of public goods. Hitherto, only “preliminary evidence” for hypothesis (1) has been provided by Simpson et al. (2012: 157), and hypothesis (2) has only been tested and confirmed by Kumru and Vesterlund (2010) and Sell (1997). Most importantly, we provide the only experimental test of the conjecture that heterogeneity in terms of status characteristics might benefit groups facing problems of collective action.

The remainder of this article is organized as follows: In the second section we give a short introduction to SCT. The third section contains a description of our experimental design and procedure as well as testable predictions in empirical terms. In the fourth section we present our experimental findings, and the fifth section discusses possible shortcomings of our approach.

Theory

Put briefly, the core argument of SCT states that whenever a group of actors faces a group task, any status characteristic that is not explicitly dissociated from the group task serves as a basis for performance expectations that in turn affect the power and prestige order of the group (cf. Figure 1)

This statement involves a number of technical concepts and demands some clarification. First of all, any task a group of actors faces qualifies as a group task as long as two scope conditions are met. That is, actors have to be oriented towards success in the task and share a collective orientation in the sense that each actor takes the other actors’ potential performance into account when deciding about her own performance level.

Secondly, the term status characteristic refers to socially significant attributes of actors with differing states, attached to which there are culturally held evaluations and expectations of competence (Correll and Ridgeway 2003: 32). For simplicity, SCT works with binary states, corresponding binary evaluations, and binary expectations of competence—for example, graduate student (positive, competent) and undergraduate student (negative, incompetent). Although all kinds of status characteristics attach evaluations to states, status characteristics differ regarding the range of activities and areas to which the expectations of competence apply. A status characteristic is termed diffuse if these expectations of competence are rather general and not restricted to specific areas. A characteristic is called specific if the expectations of competence are limited to specific areas of expertise. For instance, in many cultures females are believed to lack competence in a wide range of areas such as driving a vehicle, solving mathematical problems, and athletic performance. Hence, gender is a diffuse status characteristic in such cultural environments. In comparison, the attribute “computer expertise” carries far less implications re-

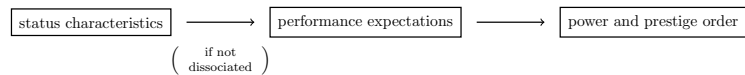


Figure 1: Graphic representation of the core argument of SCT.

garding the areas of competence and hence is an example for a specific status characteristic in most cultures (Correll and Ridgeway 2003: 32).

Thirdly, given the fact that actors want to succeed in the group task and take the behavior of others into account, they search for information regarding the relative competence in performing the group task and also the probable actions of the involved actors and organize them coherently into performance expectations.

Fourthly, the term power and prestige order refers to an important stylized fact that emerged in early research on small groups (cf. Bales 1953; Bales and Slater 1955). In small groups working on a task, the chances to perform, the initiatives to perform, the performance output, as well as the evaluations and rewards for performance are generally highly intercorrelated.

SCT is more concerned with the first arrow in Figure 1, which depicts the causal effect of status characteristics on performance expectations, than with the second arrow, which refers to the causal effect of expectations on behavioral outcomes. Practitioners of SCT (cf. Berger et al. 1977; Simpson and Walker 2002; Simpson et al. 2012) use involved graph-theoretic representations and additional theoretical concepts such as generalized expectation states or abstract task ability to study the details of the causal effect of statuses on expectations (see the online supplement). In contrast, SCT takes a quite simplistic standpoint regarding the second part of the causal chain depicted in Figure 1 (Correll and Ridgeway 2003: 31): “Once developed, performance expectation states (hereafter ‘performance expectations’) shape behavior in a self-fulfilling fashion.”

Irrespective of the technical subtleties of the graph-theoretic account of SCT, the causal chain depicted in Figure 1 implicates that actors who carry positively evaluated states of a diffuse or specific status characteristic obtain a higher position in the power and prestige order, and hence, among other things, show a greater initiative to perform and a higher output in performance towards completion of the group task. Importantly, SCT only theorizes on one particular mechanism by which status characteristics influence behavior in group tasks. That is, SCT is concerned with pure effects of expectation but does not rule out that status characteristics might influence behavior in group tasks via other mechanisms than the formation of performance expectations (cf. Driskell and Mullen 1990; Berger et al. 1977).

Note that perhaps the most interesting aspect of SCT is the fact that the theory predicts effects of statuses on behavior in group tasks that *prima facie* seem completely unrelated to any instrumental task ability involved in performing the task. That is, as long as a status is (as part of the presentation of the group task) not explicitly dissociated from the group task, actors will use this status as a basis for forming performance expectations.

Experimental Design and Procedure

Description

We conducted 12 experimental sessions in spring 2015 for which we recruited a total number of $N = 176$ subjects. Participants were enrolled as university students and recruited through the internet recruitment tool hroot (Bock, Nicklisch, and Baetge 2012). Each session lasted for about one hour and average experimental earnings amounted to 18.93 Euros, which is about twice as much as the average hourly wage for a student in this region. Each session was identical except for treatment conditions. As already indicated, there were three treatment conditions: random ($N = 60$), specific ($N = 60$), and diffuse ($N = 56$). What follows is an overview of the course of the sessions; details of each specific point can be found below.

Once all of the participants arrived, they were asked to complete a questionnaire that among other things measured their respective subjective social status. This measure was used as the diffuse status characteristic in the respective treatment. After completing the questionnaire, subjects took part in a quiz that (allegedly) served later as the specific status characteristic in the respective treatment. Then status groups were formed depending on the treatment conditions. Afterwards, subjects played the volunteer's timing dilemma. A final questionnaire completed the experiment before subjects were paid their experimental earnings.

Beginning of a session. At the beginning of the experiment, subjects were randomly seated at computer terminals within the lab.¹ Because prospect theory (Kahneman and Tversky 1979, 1992) as well as ample experimental evidence (cf. Kahneman, Knetsch, and Thaler 1990) suggest that people generally suffer more from losing a certain amount of money than they enjoy winning the very same amount, we used this so-called endowment effect (Thaler 1980) in order to generate greater incentives with a fixed monetary budget. We therefore endowed subjects with experimental money that they partially lost during the course of the experiment. In order to reinforce this effect, we informed subjects at the start of the experiment that they had earned an amount of 26 Euros, congratulated them for it, and encouraged them to think about what they could buy with that money. However, subjects were also informed that they would probably lose some of the money during the course of the experiment, depending on the decisions made by themselves and by others.

The questionnaire included several questions regarding demographics, such as gender and family situation, as well as questions regarding their educational background and course of study. To obtain the subjective social status of the students, it also included a modified version of the MacArthur Scale of subjective social status (cf. Adler et al. 2000). In contrast to the original MacArthur Scale, which mainly focuses on subjects who are already active on the labor market, our version puts more weight on typical determinants of students' social status, such as network of friends and career prospects. As in the original version, the scale consisted of a 10-rung ladder on which subjects were asked to place themselves. It was accompanied by the following instructions:

Think of this ladder as representing where students stand in Germany. At the top of the ladder are the students who are the best off—those who study the most desired academic subjects at the most respectable universities, obtain the best grades, have the best career prospects, and who are integrated into an attractive network of friends and acquaintances. At the bottom are the students who are the worst off—those who study the least desired academic subjects at the least respectable universities, obtain the worst grades, have no or dim career prospects, and who are isolated or integrated into an unattractive network of friends and acquaintances. The higher you are on this ladder, the closer you are to the students at the very top; the lower you are, the closer you are to the students at the very bottom. Where would you place yourself on this ladder?

Note that this measure of subjective status was used as a diffuse status characteristic in one of our treatments.

Quiz. The quiz was presented to the subjects as a “test of analytic skills in decision situations” and consisted of 15 questions concerning basic game-theoretical concepts such as dominance or best-response (see the online supplement for more details). These concepts were explained via examples in advance. Subjects were told that “the concepts presented are of highly practical use for yourself and others” and “your answers to these questions are pivotal for the future course of the study.” The aspired benefit of using these statements is twofold: On the one hand, these statements are supposed to underline the importance of the quiz and hence increase the legitimacy of using the quiz to form status groups in the respective treatment. On the other hand, we wanted to amplify the illusion that the points that were credited after the quiz and the number of correct answers were identical. In fact, points achieved in this quiz were totally random and had no relation to the number of correct answers.² Note that quiz points serve as a specific status characteristic in one of our treatments.

Status assignment ceremony. Following the quiz, we conducted the status assignment ceremony. Subjects were split up into two groups of equal number, either called stars (high status) or nonstars (low status). The basis for the split was publicly announced. In the random treatment, each subject drew an envelope containing a note stating that the participant was either a star or a nonstar. In the diffuse treatment, subjects were split according to their self-assessment on the modified MacArthur Scale. Therefore, subjects holding a subjective status that was above the median were assigned to the group of stars, whereas subjects that were below the median were declared to be nonstars. Subjects whose self-assessment equaled the median were randomly distributed to either group (in a manner that both groups would be equally sized). In the specific treatment, the points achieved in the quiz were used to determine group membership while following the same principles as in the diffuse treatment.

Once every subject had been informed of which group they belonged to, it was announced that the stars would be seated in a more comfortable place and would be served free soft drinks and chocolates, whereas nonstars would stay at their original places and get nothing at all. Stars and nonstars were then asked to temporarily

leave the lab, with stars being the first ones to leave. Outside the lab, both groups received written instructions concerning the volunteer's timing dilemma³ and were, depending on status group membership, either allowed to take a seat (stars) or forced to stand (nonstars). Meanwhile, inside the lab the places were rearranged so that stars would sit on one side of the lab where cabins were of a better material and more spacious as compared to those of nonstars. After the rearrangement, stars (and after them nonstars) were welcomed inside the lab and guided to their new places. As soon as all subjects were seated, drinks and chocolate were offered to the stars.

Volunteer's timing dilemma. Before starting the volunteer's timing dilemma (henceforth: VTD), subjects were asked to answer five control questions on the VTD at their computer terminal. After having selected an answer, the correct answer and an explanation were displayed to the subject, irrespective of whether the subject answered the question correctly or incorrectly. At this point, subjects were encouraged to ask questions at any time. Once the round of control questioning was completed and all upcoming questions were answered, the VTD began.

Subjects played one practice round followed by 15 regular rounds.⁴ At the beginning of each round, groups composed of two stars and two nonstars were randomly assigned anew. Importantly, each subject was aware that her group was composed of two stars and two nonstars. Then a timer started counting down from 60 seconds and caused all subjects to lose 1 cent per second until either the timer reached 0 and all subjects lost an additional 40 cents or until at least two subjects voluntarily bore a private cost of 40 cents each in order to stop the timer and end the round earlier.⁵ Over the course of a round, subjects were able to choose between two options by pressing one of two buttons labeled "Option A" and "Option B," respectively. Whereas choosing option A was equivalent to volunteering, choosing option B corresponded to not volunteering (that is, free riding). Once a subject had chosen an option, this choice was final and could not be reversed. If a subject did not choose any option before the timer had reached 0, option B was automatically chosen. During a round, subjects were completely unaware of the actions of their group members, but once the round was completed (i.e., as soon as two subjects contributed or the timer passed the 60-second mark), they were informed about the other group members' choices and respective payoffs. Because subjects were unaware of the actions of their group members during the actual play of a round and because of the fact that groups were randomly assigned anew each round, our setup does not allow for a test of the third hypothesis by Simpson et al. (2012). A formal display of the payoff function can be found in the online supplement.

Note that each group playing the VTD was composed of two stars and two nonstars. At the same time, two players were required to contribute towards the provision of the public good. The particular design of our study aimed at creating a test scenario, which ensures that our findings cannot be reduced to a pure focal point effect. The latter term refers to the stylized fact that in games involving some form of coordination, any asymmetry among the players may serve as a basis for the formation of expectations and hence behavior. Diekmann and Przepiorka (2016) demonstrated strong focal point effects in the volunteer's dilemma. This suggests that in a setup in which only one contribution is required and one star as well as

three nonstars are involved, we would expect that the star contributes simply due to a focal point effect. Note that in our design, there are two focal points (i.e., both stars or both nonstars contribute). Whereas the focal point argument provides no ground for choosing one outcome over the other, SCT supplies the sharp prediction that the stars contribute.

End of a session. Finally, each subject filled out a second questionnaire, which included a standard MacArthur Scale regarding the social status of the students' family as well as items on the educational level of the subject's parents, general happiness, religious beliefs, attitudes concerning social inequality, and prior knowledge in game theory. Once all the participants completed the questionnaire, they were paid, thanked, and dismissed.

Hypotheses. As already indicated, Simpson et al. (2012) establish that SCT applies to problems involving the voluntary provision of public goods. That is, essential scope conditions of SCT are fulfilled in classic problems of collective action such as n-person prisoner dilemmas (e.g., Hardin 1971; Hamburger 1973), linear public good games (e.g., Croson 2007; Chaudhuri 2011), or volunteer's dilemmas (cf. Diekmann 1985; Przepiorka and Diekmann 2013; Diekmann and Przepiorka 2016). In this kind of interaction situation, it is very reasonable to assume that actors are indeed oriented towards success in generating enough supply of the collective good and that actors indeed take the contributions of other actors into account while reasoning about their own level of contribution.

Because SCT applies to problems of collective action, the theoretical core argument depicted in Figure 1 supplies, among other things, two testable implications: (1) High-status actors take more initiative in contributing towards the provision of public goods than low-status actors. (2) High-status actors contribute more towards the provision of public goods than low-status actors.⁶ In addition, Simpson et al. (2012) conjecture that heterogeneity in terms of status characteristics might benefit groups facing problems of collective action because it helps to overcome startup and free-riding problems. Keep in mind that these predictions only apply to situations in which the status characteristic under consideration is not explicitly dissociated from the group task.

For clarity and future reference, we restate these hypotheses in the context of our experimental setting. The first of our central dependent variables will be called "contribution rate." This variable measures the proportion of subjects relative to the respective status group who de facto contribute (i.e., choose option A). Note that this variable somehow mixes willingness to contribute and initiative to contribute, as subjects who would be willing to contribute but wait too long (i.e., until two other team members contribute) are counted as noncontributors. Still, this variable is very close to overall performance output. Hence, hypothesis (2) by Simpson et al. (2012) translates into the following:

Individual Contribution Hypothesis (ICH):

Stars show a higher rate of contribution than nonstars in the specific and the diffuse treatment.

Because being a star or a nonstar is random in the random treatment and subjects were aware of this fact (i.e., the status characteristic is explicitly dissociated from

the group task), no performance expectations should be attached to either being a star or a no-star. We therefore state the following hypothesis:

Dissociated Contribution Hypothesis (DCH):

Rates of contribution do not differ between stars and nonstars in the random treatment.

Besides contribution rates, we will also study the willingness to contribute at any point over the course of the round in order to operationalize the initiative to contribute (i.e., hypothesis [1] by Simpson et al. 2012). That is, we will look at the pace of contribution:

Individual Initiative Hypothesis (IIH):

Stars contribute faster than nonstars in the specific and the diffuse treatment.

Analogously to DCH, we will formulate a hypothesis regarding initiative in the random treatment:

Dissociated Initiative Hypothesis (DIH):

Stars and nonstars contribute with the same pace in the random treatment.

Finally, the conjecture by Simpson et al. (2012) that status-differentiated groups are more productive will be put to a test. To do so, we will be looking at a measure of efficiency and a measure of effectiveness. First, the average loss of experimental endowments because of lag in production of the public good will be used as a measure of inefficiency. Second, the rate of successfully generated public goods (i.e., the proportion of groups in which at least two people contribute) will be considered as a measure of effectiveness. This provides two final hypotheses:

Group Efficiency Hypothesis (GEcyH):

Average loss of experimental endowments is lower in the specific and the diffuse treatment than in the random treatment.

Group Effectiveness Hypothesis (GEssH):

The rates of successfully provided public goods are higher in the specific and the diffuse treatment than in the random treatment.

Results

Figure 2 shows contribution rates by status group and treatment. We observe that contribution rates between status groups do not differ significantly in the random condition ($p < 0.538$),⁷ which confirms DCH. In stark contrast, the contribution rates of stars significantly exceed the contribution rates of nonstars in both the diffuse as well as in the specific treatment ($p < 0.002$ and $p < 0.018$, respectively), confirming ICH. Interestingly, these differences among treatments can be mainly attributed to an increased contribution rate of stars in both the diffuse and the

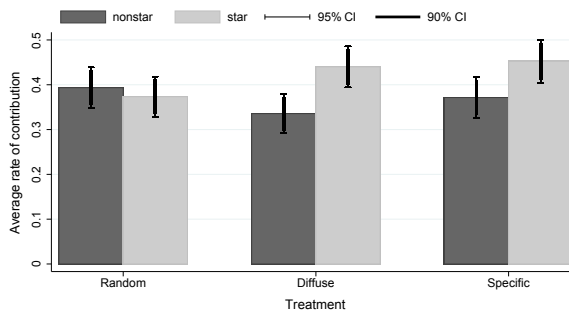


Figure 2: Mean rates of contribution by status group and treatment.

specific treatment compared to the contribution rate of stars in the random condition ($p < 0.043$ and $p < 0.019$, respectively). Although the contribution rate of nonstars differs weakly significantly between the random and diffuse treatment ($p < 0.072$), it does not differ between the random and specific treatment ($p < 0.507$). Note that neither the contribution rates of stars nor the contribution rates of nonstars differ between the diffuse and specific treatment ($p < 0.715$ and $p < 0.270$, respectively).

In order to investigate how the pace of contributions is affected by treatment and status group, we estimated three Kaplan–Meier survival models (see Figure 3). These plots measure the fraction of participants who refrain from contributing over the course of a round. At the beginning of a round (i.e., when the analysis time equals 0), the fraction of subjects who are not contributing equals 1; over the course of a round, some subjects choose to contribute and the fraction decreases. With an increase in the pace of contribution comes a faster decrease in the fraction of noncontributing participants, and therefore the curve drops earlier.⁸ We observe that descriptively, stars contribute slower than nonstars in the random treatment, although this difference is not significant ($p < 0.160$, log-rank test), which confirms DIH. Contrary to this, stars contribute significantly faster than nonstars in the diffuse treatment ($p < 0.001$, log-rank test). Similarly, in the specific treatment, the contribution pace of stars exceeds the contribution pace of nonstars, with weak significance ($p < 0.089$, log-rank test). These findings support IIIH. Finally, stars in the diffuse treatment do not show significantly more initiative in contributing towards the public good than stars in the specific treatment; however, nonstars are somewhat more reluctant in the specific treatment ($p < 0.935$ and $p < 0.075$, respectively).

We now turn to the heterogeneity conjecture formulated by Simpson et al. (2012). First, we observe that subjects lose on average approximately 68 cents more in the random treatment than in both other treatments ($p < 0.022$, t-test), supporting GEcyH (see Figure 4 [a]). Also, as visualized in Figure 4 (b), the public good is produced less often in the random treatment than in both other treatments, and this descriptive finding reaches weak statistical significance ($p < 0.088$), which confirms GEssH.

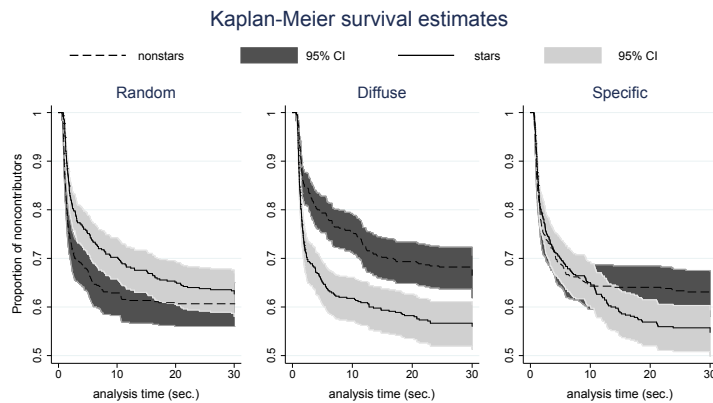


Figure 3: Kaplan–Meier survival estimates of noncontribution by status group and treatment (failure = contribution).

Finally, in order to summarize and consolidate our findings, we estimate six multivariate regressions (see Table 1). In models 1 through 3 we estimate the probability that a subject contributes with three random effects logit regressions.⁹ In models 4 through 6, survival time (i.e., the number of seconds from the start of the round until a subject contributes) is estimated by means of three random effects linear regressions.¹⁰ Models 1 and 4 are the most basic and depict the treatment effect and the effect of being a star in either the random or the nonrandom treatment. Note that this specification and in particular the inclusion of the aggregated treatment effect (variable “Nonrandom treatment?”) allows for a clean statistical test of ICH and IIH (variable “Star in nonrandom treatment?”) as well as DCH and DIH (variable “Star in random treatment?”). Models 2 and 5 add our two intrinsic status characteristics as explanatory variables—that is, subjective social status (as measured via the modified MacArthur Scale) and performance in the quiz (as measured by the number of correctly answered questions). In models 3 and 6 we add control variables. On the one hand, these included the standard demographic variables gender, age, and net income (measured via an ordinal scale).¹¹ On the other hand, we include variables that turned out to have a significant influence on the dependent variables in bivariate correlations.¹²

The variables we chose to include were two dummy variables, one measuring whether the subject had stated that she was trying to maximize her payoff and the other one measuring whether the subject held an academic degree. Adding to that we included a variable depicting the number of semesters that the subject had already been studying for. Unless otherwise specified, independent variables are dichotomous, with the value 0 for no and the value 1 for yes.

Most importantly, we find that stars in a nonrandom treatment have a significantly higher probability to contribute than nonstars in a nonrandom treatment, which confirms ICH. Similarly, they tend to contribute faster towards the public

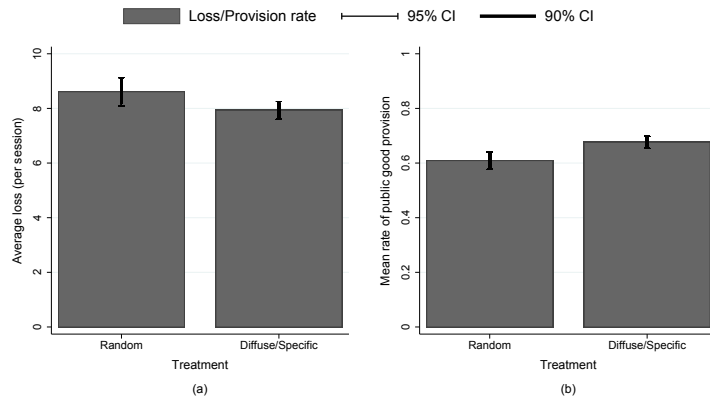


Figure 4: Average loss and mean rate of public good provision by treatment.

good, confirming IIIH. Also, we do not find any behavioral differences between stars and nonstars in the random treatment, which corroborates DCH and DIH. In addition, aggregating over both status groups, we find that neither the probability of contribution nor the pace of cooperation differ between the random and the nonrandom treatment in our study. Please note that none of our hypotheses suggest such an aggregated treatment effect. Notably, the diffuse status characteristic (i.e., subjective social status as measured via the modified MacArthur Scale) has no intrinsic influence on performance besides affecting performance expectations, as models 2 through 3 as well as models 5 through 6 corroborate. Neither the probability to contribute nor the survival time are affected by ratings on the modified MacArthur Scale. Also, the performance in the quiz has no significant influence on any of the two dependent variables.

Models 3 and 6 display that neither gender nor net income have any considerable influence on the contribution rate and pace. One variable that does have a statistically significant influence is age: with an increase in age, the likelihood of contribution increases and the pace of contribution speeds up. Further, we observe that students with more semesters of study and students who already have an academic degree tend to contribute significantly less often. Finally, those subjects who describe themselves as eager to maximize their own material payoffs show a tendency towards free-riding.

Discussion

The article at hand presents experimental evidence on how status characteristics influence performance expectations and hence performance outputs in problems involving the provision of public goods. Our results are clear-cut: we find that positively evaluated status groups contribute more and with a faster pace than

Table 1: Random effects logit (for contribution probability) and random effects linear (for contribution pace) regressions.

	<i>contribution rate (odds ratios)</i>			<i>contribution pace (in seconds)</i>		
	(1)	(2)	(3)	(4)	(5)	(6)
Nonrandom treatment?	0.782 (0.225)	0.777 (0.226)	0.640 (0.177)	3.178 (3.142)	3.175 (3.189)	5.156* (3.058)
Star in random treatment?	0.890 (0.293)	0.855 (0.280)	0.664 (0.211)	2.343 (3.607)	2.796 (3.607)	5.349 (3.542)
Star in nonrandom treatment?	1.668 [†] (0.396)	1.675 [†]	1.674 [†] (0.380)	-5.228 [†] (2.594)	-5.127* (2.684)	-4.956 [†] (2.511)
Subjective social status (1–10)		0.965 (0.070)	0.968 (0.067)		0.214 (0.800)	0.234 (0.768)
Correct quiz answers (0–15)		0.920 (0.049)	0.928 (0.049)		1.021* (0.590)	0.890 (0.591)
Male?			1.042 (0.205)			-0.215 (2.182)
Net income (1–6)			0.965 (0.105)			0.299 (1.208)
Age (years)			1.122 [†] (0.035)			-1.151 [†] (0.346)
Semesters of study			0.895 [†] (0.032)			1.139 [†] (0.387)
Academic degree?			0.326 [†] (0.073)			12.299 [†] (2.453)
Try to maximize payoff?			0.465 [†] (0.133)			9.034 [†] (3.187)
Constant	0.585 [†] (0.136)	1.613 (1.082)	0.730 (0.734)	38.08 [†] (2.551)	27.30 [†] (7.446)	36.62 [†] (11.035)
Observations	2,640	2,640	2,535	2,640	2,640	2,535

Note: standard errors in parentheses.

* $p < 0.1$; [†] $p < 0.05$.

negatively evaluated status groups in volunteer's timing dilemmas. This supports two of the most important implications of applying SCT to problems of collective action. We also find positive evidence for an interesting conjecture by Simpson et al. (2012), according to which status differentials might have a positive effect on group success in problems of collective action. In our setting, this holds true with respect to experimental earnings as well as the proportion of groups succeeding in producing the public good.

Compared to previous experimental research on the workings of status characteristics in collective action problems, one distinct advantage of our design lies in the comparison of the effects of diffuse and specific status characteristics within a fixed experimental protocol. Prima facie, Figures 2 and 3 suggest that descriptively, status groups show more pronounced behavioral differences if based on a diffuse

than on a specific characteristic. However, closer inspection revealed that in terms of statistical significance, the source of status seems to have a very subtle effect. Comparing status groups across treatments, we found that neither the rate nor the pace of contributions differ between the diffuse and the specific treatment, with the sole exception that nonstars show somewhat less initiative in contributing towards the public good if this status derives from a specific characteristic than if the status stems from a diffuse characteristic. This finding suggests that the current practice of working with either diffuse or specific status characteristics in experimental research on SCT is rather unproblematic. On a side note, this finding also sheds light on a longstanding controversy regarding the relative strength of effects of diffuse and specific statuses on the power and prestige order. Whereas the original SCT states that diffuse and specific characteristics should carry equal weight in determining the power and prestige order, Simpson and Walker (2002) argued that a conceptually more consistent reformulation of SCT lends support to the prediction that diffuse characteristics have a greater impact than specific characteristics. Because we observe a subtle yet definitive greater impact of diffuse characteristics on the pace of contributions, our results favor the reformulated version by Simpson and Walker (2002).

A second advantage of our design is that it provides convincing evidence on the mechanism by which status characteristics influence performance outputs. That is, status characteristics influence performance outputs via forming the basis of performance expectations. Three observations from our study support this claim. First, our specific treatment documents pure effects of performance expectations by separating assignment to status groups from any other property of the actors by randomization. Hence, all observed differences in performance outputs between the status groups are purely due to differential performance expectations. Second, in the random treatment we do not observe any differences in performance outputs, which is very well in line with SCT because being a star or a nonstar is explicitly dissociated from the group task in this treatment, and hence status provides no basis for performance expectations. Finally, although our observations in the diffuse treatment are more questionable on methodological grounds, they still are informative regarding pure effects of differential performance expectations. Methodologically, in the diffuse treatment we encounter the problem that any property of the subjects that is correlated with the modified MacArthur Scale of subjective social status might bias the effects of performance expectations. However, as models 2 and 3 as well as 5 and 6 indicate, subjective social status and observable behavior in the VTD are de facto uncorrelated in our data (see additional analyses in the online supplement).

Though our results speak to the power of status characteristics in influencing performance outputs in group tasks, there is an important caveat in generalizing the results. That is, our design used an extensive award ceremony and differential treatment of subjects (chocolate, lemonades, etc.) when forming the status groups. In terms of SCT, this procedure has a twofold effect. First, it ensures that the status characteristics that serve as the basis of discrimination between status groups become salient. And second, being a star as well as getting drinks and chocolate is intrinsically rewarding. As argued by an extension of SCT to reward expectations

and structures (Berger et al. 1985; Webster and Hysom 1998), actors generally expect that high-status actors obtain greater rewards than low-status actors. Hence, providing these benefits to stars but not to nonstars should theoretically bolster the differential performance expectations; Hysom (2009) corroborates this reasoning with experimental evidence. Of course, we implemented this feature of our design to ensure that subjects got the impression that at least during the duration of the experiment, being a star or a nonstar had some meaning besides being a mere label displayed on a computer screen. Because this feature was present in all three experimental conditions, we have no way of isolating or estimating its effects. Hence, this leaves open the question of to what extent our results generalize to status characteristics outside the lab that come without such award ceremonies and differential treatment. Although this marks a limitation of our study, it has to be kept in mind that many established status characteristics such as gender, ethnicity, and age do de facto come with differential treatment and rewards in our societies. For instance, wages and income (i.e., the most important form of material rewards) generally rise with age (cf. Hellerstein, Neumark, and Troske 1999; Hall and Farkas 2008), and discrimination of members of ethnic minorities (cf. Pager, Bonikowski, and Western 2009; Blommaert, Coenders, and van Tubergen 2013; Bursell 2014) as well as the gender pay gap (cf. Aisenbrey and Brückner 2008; Auspurg, Hinz, and Sauer 2017) are realities on labor markets. In addition, the differential treatment of status groups is often part of our everyday culture and subject to strong and sanctioned norms of conduct. To cite an illuminating facet, studies have shown that teachers judge as well as respond differently to performance outputs by boys and girls during class (cf. Jones and Wheatley 1990; Tiedemann 2002). In this sense, the award ceremony and differential treatment of our subjects just mimics fundamental aspects of social reality.

Although the results of existing experimental research look promising regarding the empirical validity of SCT applied to problems of collective action, the data basis still is rather slim and the question of to what extent these findings are reliable remains. Hence, future research on the application of SCT to problems of collective action should address the question of whether these observations are robust regarding variations in essential parameters of our design. These variations should cover different types of public goods (i.e., different “technologies” that translate inputs into outputs; cf. Sandler 1992) as well as different diffuse and specific status characteristics.

Notes

- 1 The whole experiment was programmed and conducted with the software z-Tree (Fischbacher 2007).
- 2 All subjects in the specific treatments were debriefed after the completion of the study.
- 3 Although the text was the same for all subjects, instructions of stars were decorated with an additional golden star.
- 4 In order to avoid endgame effects, subjects were not informed about how many rounds they would play.

- 5 Strictly speaking, in a volunteer's dilemma it is only necessary for one participant to cooperate. In our experiment we use a generalization of the volunteer's dilemma in which two volunteers are required. The volunteer's dilemma is a special case of a step-level public good game. The characterizing feature of a step-level public good is that it is provided if and only if a certain threshold of contributions towards the public good is met. Failure in meeting the threshold results in no or an inefficient provision of the public good. Also, contributions in excess of the threshold are inefficient. Real-world examples for step-level goods are, for instance, fundraising for community projects such as building a bridge or a fence (cf. Andreoni 1998).
- 6 The reader might wonder why we are interested in the initiative to contribute on top of total contributions. This interest is related to the fact that in many real-life public good problems, there is a certain time limit for the provision of the public good or the value of the public good degenerates over time. For instance, if a person is in urgent need of medical help because of an emergency, her condition might deteriorate until help is provided, with possibly irreversible effects. Note that the VTD models a situation in which the value of contributions towards the provision of the public good shrinks over time. Additionally, Simpson et al. (2012) relates the initiative to contribute to the so-called "start-up problem" in collective action. That is, many collective action problems are characterized by a "critical mass" incentive structure, in which it is beneficial to contribute towards the provision of the public good provided that enough contributions by other actors are made (Oliver, Marwell, and Teixeira 1985). In situations like these, the initiative to make contributions is vital for the provision of public goods because initial contributors might trigger contributions by other actors.
- 7 Unless otherwise specified, χ^2 tests are used.
- 8 Because noncontribution rates only change very marginally after 30 seconds, the plots are truncated, and the final drops at 30 seconds represent the change of the noncontribution rates at the end of the whole round (i.e., after 60 seconds).
- 9 Additionally, we ran clustered logit regressions, which yielded the same results.
- 10 Additionally, we ran clustered Cox regressions as well as several random effects parametric survival models; these models provided the very same results (including exponential, lognormal, and Weibull survival distribution). Also, using logarithmic survival times as the dependent variable in models 4 through 6 does not affect our findings.
- 11 Specifically, net income was measured in steps of 300 Euros, i.e., 0 to 300, 301 to 600, etc.
- 12 Because we lack theoretical reasons to include specific control variables, we opted to approach the task of selecting control variables on purely empirical grounds. Note that because of randomization, controlling for confounds is pointless with regard to the research interest of this article. Still, the reported effects might be interesting per se.

References

- Adler, Nancy E., Elissa S. Epel, Grace Castellazzo, and Jeannette R. Ickovics. 2000. "Relationship of Subjective and Objective Social Status with Psychological and Physiological Functioning: Preliminary Data in Healthy, White Women." *Health Psychology* 19:586–92. <https://doi.org/10.1037/0278-6133.19.6.586>.
- Aisenbrey, Silke, and Hannah Brückner. 2008. "Occupational Aspirations and the Gender Gap in Wages." *European Sociological Review* 24:633–49. <https://doi.org/10.1093/esr/jcn024>.

- Andreoni, James. 1998. "Toward a Theory of Charitable Fund-Raising." *Journal of Political Economy* 106:1186–213. <https://doi.org/10.1086/250044>.
- Auspurg, Katrin, Thomas Hinz, and Carsten Sauer. 2017. "Why Should Women Get Less? Evidence on the Gender Pay Gap from Multifactorial Survey Experiments." *American Sociological Review* 82:179–210. <https://doi.org/10.1177/0003122416683393>.
- Bales, Robert F. 1953. "The Equilibrium Problem in Small Groups." In *Working Papers in the Theory of Action*, edited by Talcott Parsons, Robert F. Bales, and E. A. Shils, pp. 111–61. Glencoe: Free Press.
- Bales, Robert F., and Philip E. Slater. 1955. "Role Differentiation in Small Decision Making Groups." In *Family, Socialization, and Interaction Process*, edited by Talcott Parsons and Robert F. Bales, pp. 259–306. Glencoe: Free Press.
- Berger, Joseph, M. Hamit Fisek, Robert Zane Norman, and Morris Zelditch, Jr. 1977. *Status Characteristics and Social Interaction: An Expectation-States Approach*. New York: Elsevier.
- Berger, Joseph, Susan J. Rosenholtz, and Morris Zelditch. 1980. "Status Organizing Processes." *Annual Review of Sociology* 6:479–508. <https://doi.org/10.1146/annurev.so.06.080180.002403>.
- Berger, Joseph, David G. Wagner, and Morris Zelditch, Jr. 1985. "Introduction: Expectation States Theory - Review and Assessment." In *Status, Rewards, and Influence: How Expectations Organize Behaviour*, edited by Joseph Berger and Morris Zelditch, Jr., pp. 1–72. San Francisco: Jossey-Bass.
- Blommaert, Lieselotte, Marcel Coenders, and Frank van Tubergen. 2013. "Discrimination of Arabic-Named Applicants in the Netherlands: An Internet-Based Field Experiment Examining Different Phases in Online Recruitment Procedures." *Social Forces* 92:957–82. <https://doi.org/10.1093/sf/sot124>.
- Bock, Olag, Andreas Nicklisch, and Ingmar Baetge. 2012. "hroot: Hamburg Recruitment and Organization Online Tool." *WiSo-HH Working Paper Series No. 1*.
- Bursell, Moa. 2014. "The Multiple Burdens of Foreign-Named Men – Evidence from a Field Experiment on Gendered Ethnic Hiring Discrimination in Sweden." *European Sociological Review* 30:399–409. <https://doi.org/10.1093/esr/jcu047>.
- Chaudhuri, Ananish. 2011. "Sustaining Cooperation in Laboratory Public Goods Experiments: A Selective Survey of the Literature." *Experimental Economics* 14:47–83. <https://doi.org/10.1007/s10683-010-9257-1>.
- Coleman, James Samuel. 1990. *Foundations of Social Theory*. Cambridge, Mass: Belknap Press.
- Correll, Shelley J., and Cecilia L. Ridgeway. 2003. "Expectation States Theory." In *Handbook of Social Psychology*, edited by John Delamater, pp. 29–53. New York: Kluwer Academic/Plenum Publishers.
- Croson, Rachel T. A. 2007. "Theories of Commitment, Altruism and Reciprocity: Evidence from Linear Public Goods Games." *Economic Inquiry* 45:199–216. <https://doi.org/10.1111/j.1465-7295.2006.00006.x>.
- Diekmann, Andreas. 1985. "Volunteer's Dilemma." *Journal of Conflict Resolution* 29:605–10. <https://doi.org/10.1177/0022002785029004003>.

- Diekmann, Andreas, and Wojtek Przepiorka. 2016. "Take One for the Team! Individual Heterogeneity and the Emergence of Latent Norms in a Volunteer's Dilemma." *Social Forces* 94:1309–333. <https://doi.org/10.1093/sf/sov107>.
- Driskell, James E., and Brian Mullen. 1990. "Status, Expectations, and Behavior: A Meta-Analytic Review and Test of the Theory." *Personality and Social Psychology Bulletin* 16:541–53. <https://doi.org/10.1177/0146167290163012>.
- Fischbacher, Urs. 2007. "z-Tree: Zurich Toolbox for Ready-made Economic Experiments." *Experimental Economics* 10:171–78. <https://doi.org/10.1007/s10683-006-9159-4>.
- Hall, Matthew, and George Farkas. 2008. "Does Human Capital Raise Earnings for Immigrants in the Low-Skill Labor Market?" *Demography* 45:619–39. <https://doi.org/10.1353/dem.0.0018>.
- Hamburger, Henry. 1973. "N-Person Prisoner's Dilemma." *Journal of Mathematical Sociology* 3:27–48. <https://doi.org/10.1080/0022250X.1973.9989822>.
- Hardin, Russell. 1971. "Collective Action as an Agreeable N-Prisoners' Dilemma." *Behavioral Science* 16:472–81.
- Heckathorn, Douglas D. 1996. "The Dynamics and Dilemmas of Collective Action." *American Sociological Review* 61:250–77. <https://doi.org/10.2307/2096334>.
- Hellerstein, Judith K., David Neumark, and Kenneth R. Troske. 1999. "Wages, Productivity, and Worker Characteristics: Evidence from Plant-Level Production Functions and Wage Equations." *Journal of Labor Economics* 17:409–46. <https://doi.org/10.1086/209926>.
- Homans, George C. 1974. *Social Behavior: Its Elementary Forms*. Oxford: Harcourt Brace Jovanovich.
- Hysom, Stuart J. 2009. "Status Valued Goal Objects and Performance Expectations." *Social Forces* 87:1623–48. <https://doi.org/10.1353/sof.0.0160>.
- Jones, M. Gail, and Jack Wheatley. 1990. "Gender Differences in Teacher-Student Interactions in Science Classrooms." *Journal of Research in Science Teaching* 27:861–74. <https://doi.org/10.1002/tea.3660270906>.
- Kahneman, Daniel, Jack L. Knetsch, and Richard H. Thaler. 1990. "Experimental Tests of the Endowment Effect and the Coase Theorem." *Journal of Political Economy* 98:1325–48. <https://doi.org/10.1086/261737>.
- Kahneman, Daniel, and Amos Tversky. 1979. "Prospect Theory: An Analysis of Decision under Risk." *Econometrica* 47:263–92. <https://doi.org/10.2307/1914185>.
- Kahneman, Daniel, and Amos Tversky. 1992. "Advances in Prospect Theory: Cumulative Representation of Uncertainty." *Journal of Risk and Uncertainty* 5:297–323. <https://doi.org/10.1007/BF00122574>.
- Kumru, Cagri S., and Lise Vesterlund. 2010. "The Effect of Status on Charitable Giving." *Journal of Public Economic Theory* 12:709–35. <https://doi.org/10.1111/j.1467-9779.2010.01471.x>.
- Liebe, Ulf, and Andreas Tutić. 2010. "Status Groups and Altruistic Behaviour in Dictator Games." *Rationality and Society* 22:353–80. <https://doi.org/10.1177/1043463110366232>.

- Mauss, Marcel. [1925] 1954. *The Gift: Forms and Functions of Exchange in Primitive Societies*. London: Cohen and West.
- Oliver, Pamela, Gerald Marwell, and Ruy Teixeira. 1985. "A Theory of the Critical Mass. I. Interdependence, Group Heterogeneity, and the Production of Collective Action." *American Journal of Sociology* 91:522–56. <https://doi.org/10.1086/228313>.
- Olson, Mancur. 1965. *The Logic of Collective Action: Public Goods and the Theory of Groups*. Harvard University Press. <https://doi.org/10.1017/CB09780511807763>.
- Ostrom, Elinor. 1990. *Governing the Commons: The Evolution of Institutions for Collective Action*. Political Economy of Institutions and Decisions. Cambridge and New York: Cambridge University Press.
- Otsubo, Hironori, and Amnon Rapoport. 2008. "Dynamic Volunteer's Dilemmas over a Finite Horizon: An Experimental Study." *Journal of Conflict Resolution* 52:961–84.
- Pager, Devah, Bart Bonikowski, and Bruce Western. 2009. "Discrimination in a Low-Wage Labor Market: A Field Experiment." *American Sociological Review* 74:777–99. <https://doi.org/10.1177/000312240907400505>.
- Piff, Paul K., Michael W. Kraus, Stéphane Côté, Bonnie Hayden Cheng, and Dacher Keltner. 2010. "Having Less, Giving More: the Influence of Social Class on Prosocial Behavior." *Journal of Personality and Social Psychology* 99:771–84. <https://doi.org/10.1037/a0020092>.
- Przepiorka, Wojtek, and Andreas Diekmann. 2013. "Individual Heterogeneity and Costly Punishment: A Volunteer's Dilemma." *Proceedings of the Royal Society B* 280:20130247. <https://doi.org/10.1098/rspb.2013.0247>.
- Sandler, Todd. 1992. *Collective Action: Theory and Applications*. Ann Arbor: University of Michigan Press.
- Sell, Jane. 1997. "Gender, Strategies, and Contributions to Public Goods." *Social Psychology Quarterly* 60:252–65. <https://doi.org/10.2307/2787085>.
- Simpson, Brent, and Henry A. Walker. 2002. "Status Characteristics and Performance Expectations: A Reformulation." *Sociological Theory* 20:24–40. <https://doi.org/10.1111/1467-9558.00149>.
- Simpson, Brent, and Robb Willer. 2015. "Beyond Altruism: Sociological Foundations of Cooperation and Prosocial Behavior." *Annual Review of Sociology* 41:43–63. <https://doi.org/10.1146/annurev-soc-073014-112242>.
- Simpson, Brent, Robb Willer, and Cecilia L. Ridgeway. 2012. "Status Hierarchies and the Organization of Collective Action." *Sociological Theory* 30:149–66. <https://doi.org/10.1177/0735275112457912>.
- Thaler, Richard H. 1980. "Toward a Positive Theory of Consumer Choice." *Journal of Economic Behavior and Organization* 1:39–60. [https://doi.org/10.1016/0167-2681\(80\)90051-7](https://doi.org/10.1016/0167-2681(80)90051-7).
- Tiedemann, Joachim. 2002. "Teachers' Gender Stereotypes as Determinants of Teacher Perceptions in Elementary School Mathematics." *Educational Studies in Mathematics* 50:49–62. <https://doi.org/10.1023/A:1020518104346>.
- Webster, Murray, and Stuart J. Hysom. 1998. "Creating Status Characteristics." *American Sociological Review* 63:351–78. <https://doi.org/10.2307/2657554>.

Weesie, Jeroen. 1993. "Asymmetry and Timing in the Volunteer's Dilemma." *Journal of Conflict Resolution* 37:569–90. <https://doi.org/10.1177/0022002793037003008>.

Acknowledgements: Financial support by the German Research Foundation (DFG TU 409/1) and research assistance by Maximilian Lutz are gratefully acknowledged.

Andreas Tutić: Institute of Sociology, Leipzig University.
E-mail: andreas.tutic@sozio.uni-leipzig.de.

Sascha Grehl: Institute of Sociology, Leipzig University.
E-mail: sascha.grehl@uni-leipzig.de.

Intuition, Reflection, and Prosociality: Evidence from A Field Experiment

Sascha Grehl, Andreas Tutić

PLoS ONE 17(2), e0262476 (2022).

RESEARCH ARTICLE

Intuition, reflection, and prosociality: Evidence from a field experiment

Sascha Grehl^{1*}, Andreas Tutić

Institut für Soziologie, Leipzig University, Leipzig, Saxony, Germany

* sascha.grehl@uni-leipzig.de

Abstract

Are humans instinctively good or is it only our capacity for reflection that enables us to restrain our selfish traits and behave prosocially? Against the background of dual-process theory, the question of whether people tend to behave prosocially on intuitive grounds has been debated controversially for several years. Central to this debate is the so-called social heuristic hypothesis (SHH), which states that subjects orient their behavior more closely to their deeply ingrained norms and attitudes when the behavior comes about in an intuitive rather than reflective manner. In this paper, we apply the SHH to a novel setting and investigate whether its implications hold true in a non-reactive field experiment, in which subjects are unaware that they are part of a study. We test whether subjects report a misdirected email or try to use the opportunity to reap a monetary benefit. Since all subjects participated six months prior to the field experiment in a lab experiment, we have solid measures of the subjects' general tendency to behave intuitively and their prosocial attitudes. In addition, participants were asked in a follow-up survey to self-report their intuitiveness at the time of the decision. While we observe a significant and positive effect on prosocial behavior for self-reported intuitiveness (but not for general intuitiveness) in the bivariate analyses, this effect becomes insignificant when controlling for interaction effects with attitudes. In addition, for both forms of intuitiveness, we find a significant and positive interaction effect with subjects' prosocial attitudes on prosocial behavior. Hence, this study confirms previous findings from laboratory as well as online studies and provides external validity by demonstrating that the SHH applies in a real-life situation.



OPEN ACCESS

Citation: Grehl S, Tutić A (2022) Intuition, reflection, and prosociality: Evidence from a field experiment. PLoS ONE 17(2): e0262476. <https://doi.org/10.1371/journal.pone.0262476>

Editor: Pablo Brañas-Garza, Universidad Loyola Andalucía Cordoba, SPAIN

Received: July 13, 2021

Accepted: December 24, 2021

Published: February 25, 2022

Copyright: © 2022 Grehl, Tutić. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the paper and its [Supporting information files](#).

Funding: AT was financially supported by the German Research Foundation (Grant number: TU 409/3-1) <https://www.dfg.de/>. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Introduction

In many everyday situations as well as in many experimental studies, it has been observed time and again that a substantial proportion of people are willing to act prosocially, i.e., to forgo short-term benefits or incur some personal costs in order to improve the well-being of others [1–3]. These findings pose explanatory challenges in situations where mechanisms such as reputation or reciprocity cannot contribute to the emergence of prosocial behavior, i.e., interactions that are anonymous and one-time.

In the context of explaining prosocial behavior in one-time interactions, dual process theory (DPT) has proven to be particularly fruitful in recent years. According to the DPT, human

cognition and action can only be explained by the interplay of two qualitatively distinct kinds of systems or types of mental processes, i.e., automatic-spontaneous (intuitive) processes are contrasted with controlled-deliberative (reflective) processes [4–6]. Automatic-spontaneous processes are typically automatic, fast, and associative, operate outside of the actor's consciousness and in parallel. In contrast, controlled-deliberative processes are typically controlled, slow, and rule-based, operate within the actor's consciousness and only in serial succession. While the exact interplay between these two kinds of processes is still a matter of debate [6, 7], DPT has nevertheless given rise to influential applications in research on prosocial behavior.

In particular, the idea that acting prosocially is an intuitive human behavior has received much attention. In short, this idea states that the more someone relies on his or her intuitiveness, i.e., the tendency to engage in an intuitive decision-making process, the more prosocially he or she will act [8, 9]. This hypothesis, which we will call the *intuitive prosociality hypothesis*, has been confirmed in many studies: First, in observational studies, where response latencies [8, 10, 11] and dispositional measures of thinking dispositions and cognitive styles [12] are used to assess subjects' intuitiveness. Second, in experimental studies which aim at directly influencing subjects' intuitiveness through procedures such as time pressure [8, 13, 14], cognitive load [15, 16], or ego-depletion [17].

At the same time, a number of skeptical contributions have appeared that question the empirical validity of the intuitive prosociality hypothesis on both theoretical as well as methodological grounds. In terms of methodology, it was argued that observational response latencies do not allow a straightforward inference about the type of cognitive process involved, because response latencies are also influenced, among other things, by the strength of preferences (discriminability of alternatives) [18, 19]. Some studies uncovered that intuition can have a positive influence on certain aspects of prosociality (e.g., egalitarian choices), while also having a negative influence on other aspects of prosociality (e.g., social efficient choices) [20, 21]. Further, it was also shown that by appropriately manipulating the payoff structure it can be observed that actors who tend to decide quickly behave less prosocially than actors who tend to decide slowly [18]. Moreover, even studies which employ experimental manipulations of the time available in decision-making do not necessarily yield results that support the intuitive prosociality hypothesis [22–24].

In light of this contradictory evidence, it became clear that the original hypothesis could not be upheld without additional qualifying statements. Thus, the *social heuristic hypothesis* (SHH) took its place [13]. The SHH consists of two parts: First, it states that humans internalize strategies that tend to be beneficial in their daily social life. These internalized strategies function as cues about how to behave in a new and unfamiliar social interaction. Second—and this is where DPT comes into play—the hypothesis states that actors who decide intuitively will follow these cues more closely than actors who reflect on the situation, thereby potentially recognizing that an alternative behavior is more advantageous.

The simplified idea behind this hypothesis is that decisions are less costly if they are made via an intuitive process rather than a reflective process. If the average cost saved by an intuitive decision (compared to the reflective decision) exceeds the average harm caused by this decision (in comparison to the reflective decision), intuitiveness may be evolutionarily stable [25–27]. Since prosocial behavior is—by definition—never advantageous in a one-time anonymous interaction, deliberation will undermine prosociality, whereas intuition might favor both cooperative as well as selfish behavior depending on prior experience. In other words, the nature of everyday social interactions is a moderating factor that influences the intuitive responses of actors. Assuming that it is true for most societies that prosocial behavior is more successful in daily life, the SHH can explain why intuitive actors might act more prosocially than their less intuitive counterparts. Furthermore, it can also explain why this might be not true for certain situations, societies, or parts thereof [12, 28].

In this paper, we focus on another potential moderator of the relationship between intuitiveness and prosocial behavior, namely prosocial attitudes [11, 29–31]. The basic idea is that prosocial attitudes, similar to previous experiences, also function as normative cues, but at the individual level. Therefore, the SHH can also be applied with respect to prosocial attitudes. Indeed, studies on this topic find that individuals with stronger prosocial attitudes act more prosocial. More interestingly, they also find evidence that prosocial individuals act less prosocial the longer they need to decide [29, 30, 32] and that similar results are observed when an experimental manipulation like time pressure is used [33].

The objective of this paper is to apply the SHH to a novel setting and investigate whether its implications hold true in a real-life situation. In doing so, we use a non-reactive field experiment in which subjects are unaware that they are part of a study. For this purpose, a *non-reactive field experiment* was conducted with subjects previously enrolled in a lab experiment at an Experimental Laboratory (LAB) [34]. The participants received an apparently misdirected email from the official email address of the lab, which contained a payoff code that allowed participants to redeem a monetary payment online. Since the code and the associated money were obviously intended for another person, we classify the attempt to redeem the code as less prosocial than simply ignoring the email. In addition, we classify both of these actions as less prosocial than the action of informing us of the alleged error. In this regard, we closely follow the psychological definition of prosociality, which refers to observed behavior, but not to underlying motivations [20, 35]. Thus, it is possible that an action we classify as prosocial might be driven by selfish rather than altruistic motives. Within the study, we experimentally varied three different parameters: First, the amount of money promised in the email, second, the verbal framing of the email, and third, the presence of a disclaimer. However, these experimental manipulations are only of secondary interest, as our main focus is on the interaction effects between attitudes and intuitiveness. For this purpose, we measured participants' intuitiveness in two different ways: First, we asked participants in a follow-up survey to recall the situation in which they decided how to respond to the email and to self-assess their intuitiveness at this moment. A second measure indicating the participants' general tendency toward intuitive behavior was collected in the previously conducted lab experiment using the so-called Cognitive Reflection Test [36]. Note that neither measurement is based on response time, which, as mentioned earlier, is often considered a problematic measure [18, 19]. In the lab experiment, we also obtained the prosocial attitude of the participants.

Following previous findings in the literature on attitudes and prosocial behavior [29, 30, 32], we expect that people with stronger prosocial attitudes will behave more prosocially. Furthermore, and in accordance with the SHH, we expect a positive interaction effect between prosocial attitudes and intuitiveness on prosocial behavior. Or, phrased in the opposite way, the negative effect of reflection (non-intuitiveness) on prosocial behavior should be most pronounced for subjects with strong prosocial attitudes and become less pronounced the weaker these attitudes are.

The remainder of the paper is organized as follows. In the next section the methods and the design of the field experiment are presented. The following section reports our empirical findings, and in the last section we draw conclusions with respect to the SHH and the DPT, discuss limitations of the current study, and outline directions for future research.

Methods

In the style of the non-reactive lost letter method [37, 38], we developed a field experiment in which we sent apparently misdirected emails to which recipients could respond in different ways. The emails were sent from the official address of the Leipzig Experimental Laboratory

(LAB) to former participants of a previous lab experiment conducted about six months before this study [34]. The field experiment was conducted in accordance with the Declaration of Helsinki and all procedures were approved by the Institute of Sociology of the Leipzig University.

The content of the emails was standardized and the alleged recipient of the email was a woman named “Marion Koch”. This name was chosen because, first, it is a relatively typical German name and, second, none of the recipients had either this first or last name, so it should be apparent to participants that they were not the intended recipients of the email. In the email, we thanked the recipients for allegedly participating in a lab experiment the day before, told them how much money they had earned in total, and sent them a payout code to receive the money anonymously (for the complete content of the email, see [S1 Table](#)). The recipients were told that the payout code had to be entered and submitted on the LAB website to be redeemed for an Amazon voucher, a PayPal credit, or a payout voucher by which the amount in cash could be anonymously received at the LAB. However, this website was prepared for the experiment in such a way that an error message appeared when the code was entered (“Unfortunately, this code has already been used.”) and the code, as well as the time of the attempt, were recorded.

For the experimental manipulation, we used a $2 \times 2 \times 2$ between-participants factorial design. We varied the temptation, i.e., the amount of money Ms. Koch was supposedly entitled to, by indicating a payout amount of 6.25€ (temptation low treatment) and 21.75€ (temptation high treatment), respectively. Furthermore, the email either contained a disclaimer at the end (disclaimer treatment) or not (no-disclaimer treatment). The disclaimer informed the participants that this email was “intended exclusively for the person addressed” and that the LAB should be contacted if this email was not delivered to the rightful recipient. Finally, we varied the description of Ms. Koch’s action that had led to the earning of the money so that either it was described as “contribute to a group fund” (contribution treatment) or “steal from a group fund” (theft treatment).

The participants could respond in three possible ways: They could simply ignore the email, notify us of our “mistake”, or attempt to use the code. For simplicity, these behaviors will be referred to as neutral, helpful, and selfish actions in the following. In fact, there were some people who showed several responses, such as reporting the error first and using the code later. In such cases, we classified the person according to which response occurred first. This was possible because we had collected the date and time for each response.

At the end of the two-week data collection period, all participants were informed about the field experiment and invited to take part in an online follow-up survey. At this point, participants had the opportunity, in accordance with the Declaration of Helsinki, to object to their participation in the study. The primary goal of this follow-up survey was to measure the intuitiveness participants exhibited when they decided on how to respond to the email. This was done using a series of questions. For example, they were asked whether they had considered other alternative actions, how quickly they had made their decision, and to what extent they based their action on gut feeling (see [S2 Table](#)).

Since only persons who had previously participated in the lab experiment took part in the field experiment, we also have further information from the lab experiment at our disposal. In particular, two variables from the lab experiment are central to this study: First, the prosocial attitude of the participants, which is measured using a short version of the Prosocial Personality Battery [39] (see [S3 Table](#)), and second, the general intuitive tendency of the participants, which is measured using an extended version of the Cognitive Reflection Test (CRT) [36, 40]. The extended version of the CRT consists of the three original questions and one additional question [41] (see [S4 Table](#)). Following [12], we use the number of wrong answers in the CRT

as a measure of a general intuitiveness tendency of the participants [20, 42]. In addition, age, gender, or lab experience serve as control variables.

Results

A total of 763 participants of a previously conducted lab experiment could be reached by email and these constitute the participants of our field experiment. Of these, 19 individuals expressed suspicion during the data collection period that the email could be part of a study and were therefore excluded from further analyses. In addition, one of these individuals later decided to drop out of the study during the follow-up interview. Thus, a total of 744 valid observations are available for analysis. Of these, 35% and 65% report being male and female, respectively. The majority of our subjects (over 80%) are students and the average age is 25 years.

With respect to our dependent variable, i.e., the behavioral response to our email, we find that about half of the participants behaved helpfully and notify us of our alleged error, approximately 40% showed no reaction at all and are therefore classified as neutral, and about 10% of the participants behaved selfishly and attempt to enter the payout code online. In total, 485 subjects (65%) participated in the follow-up survey, with participation rates above 50% in each of the three response groups (Fig 1); however, three subjects did not answer all questions regarding the intuitiveness of their response, so only 482 complete cases are available.

Let us now turn to the central independent variables of the field experiment. As a measure of prosociality, we use an index based on the short version of the Prosocial Personality Battery. Therefore, we selected 13 questions of the Prosocial Personality Battery which maximize reliability (Cronbach's $\alpha = .80$) while minimizing the number of factors in a factor analysis (the selected questions can be found in S3 Table). This index can take values from 0 (= low prosocial attitudes) to 1 (= high prosocial attitudes) and will be referred to as the prosocial attitude (PSA) score. Fig 2(a) shows the distribution of the PSA score: on average, participants have a value of .677 with a standard deviation of .108. Less than 5% of the participants have a PSA

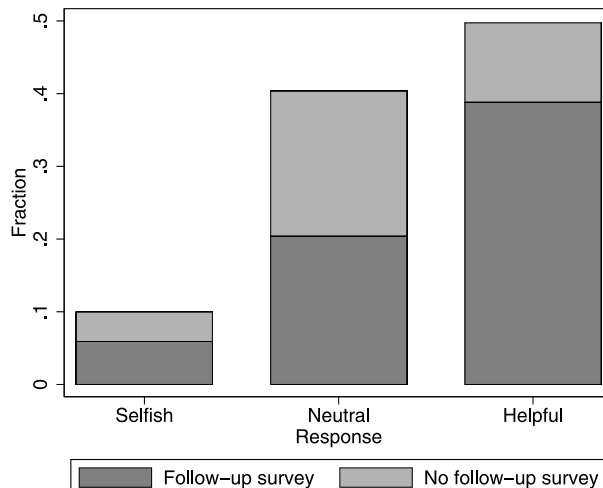


Fig 1. Overview response and participation. Distribution of participants' responses in the field experiment and proportion of those who participated in the follow-up survey.

<https://doi.org/10.1371/journal.pone.0262476.g001>

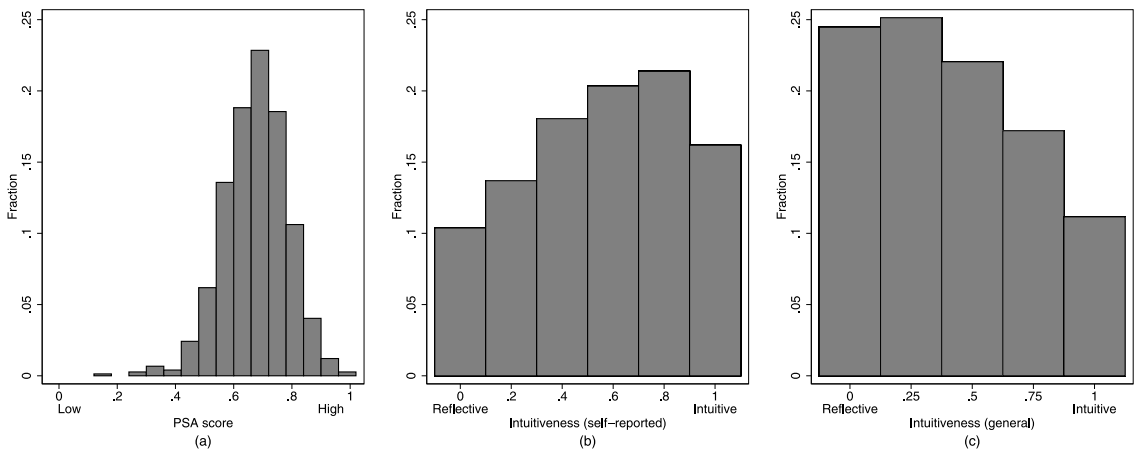


Fig 2. Overview main independent variables. Distribution of participants' (a) PSA score, (b) self-reported intuitiveness, and (c) general intuitiveness.

<https://doi.org/10.1371/journal.pone.0262476.g002>

score lower than 0.5, which means that we are dealing with a relatively prosocial sample in which there are hardly any participants with low prosocial attitudes. Therefore, if we were to compare two groups for the sake of illustration, it would be those with intermediate and high prosocial attitudes.

Regarding self-reported intuitiveness, we use an index calculated from the questions of the followup-survey. The index can take values from 0 (reflective) to 1 (intuitive) and has acceptable reliability with a Cronbach's α of .70. The distribution of self-reported intuitiveness is a slightly left-skewed normal distribution (see Fig 2(b)). Furthermore, in Fig 2(c) we see a right-skewed distribution of the general tendency toward intuitive behavior (in the following: general intuitiveness), which is calculated using the number of wrong answers in the extended CRT. This measure can take values between 0 (reflective) and 1 (intuitive) and has a Cronbach's α of .64.

Regarding possible correlations between our main independent variables, we find that the correlation between PSA score and general intuitiveness is not significant ($r = .027, p = .460$), while we observe a significant positive correlation between PSA score and self-reported intuitiveness ($r = .106, p = .020$). Hence, participants with higher prosocial attitudes are slightly more intuitive. Interestingly, the correlation between general intuitiveness and self-reported intuitiveness is negligibly weak ($r = .039, p = .397$).

In our analyses, we use participants' first response in our field experiment as the dependent variable. The observed reactions can be ranked with respect to prosociality: Redeeming the code is the least prosocial action, followed by the neutral action and then by the helpful action of reporting the error. Thus, we use ordered-logit models (OLMs) to test our hypotheses. In these models, a positive regression coefficient of a variable indicates that an increase in that variable increases the probability of a more prosocial action, i.e., instead of the selfish action, the neutral or helpful action is chosen, or instead of the neutral action, the helpful action is chosen. Similarly, a negative coefficient indicates that an increase in this variable decreases the likelihood of a prosocial action. However, an important prerequisite for using ordered logit models is the proportional odds assumption, which states that the calculated odds ratios must be the same for each of the ordered dichotomizations of the dependent variable [43]. To check

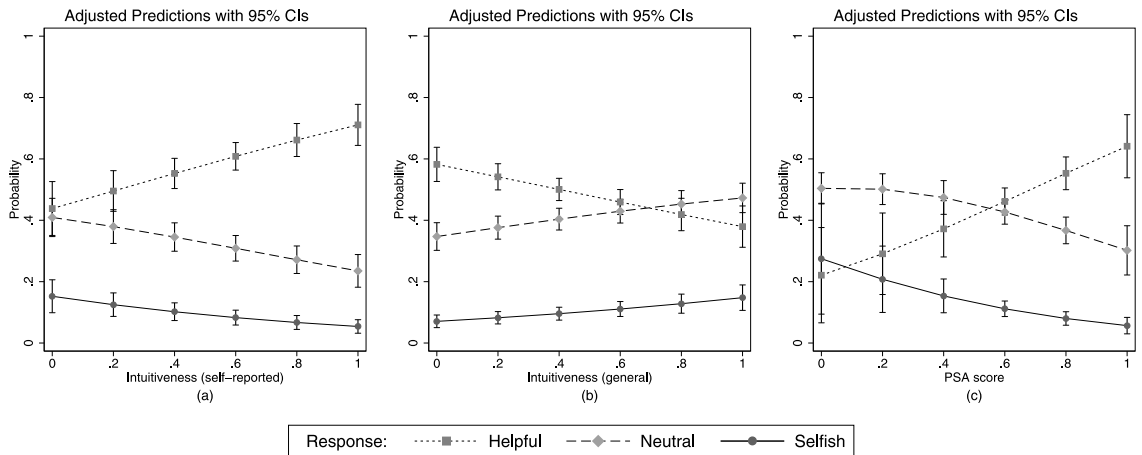


Fig 3. Bivariate analyses. Predicted probabilities of behavioral responses as a function of (a) self-reported intuitiveness, (b) general intuitiveness, and (c) PSA score.

<https://doi.org/10.1371/journal.pone.0262476.g003>

whether this assumption is warranted, we used the Brant test [44]. It turned out that none of our models violates this assumption. As a test of robustness, we performed analogous analyses using OLS models. All variables which are significant under the OLM specification prove to be significant under the OLS specification as well.

Consistent with the intuitive prosociality hypothesis, in the bivariate analysis (see Fig 3(a)) we find that a significant positive effect on prosociality can be observed for self-reported intuitiveness ($\beta = 1.148, p < .001$). However, the opposite is true for the general tendency towards intuitive behavior (Fig 3(b)); we observe that individuals tend to be less prosocial the higher their general intuitiveness ($\beta = -.824, p < .001$). Finally, as previous studies have shown [29, 33], we observe a significant and positive effect of the prosocial attitude on prosocial behavior ($\beta = 1.841, p = .006$, Fig 3(c)).

To test the interaction effects between intuitiveness and the PSA score as posited by the SHH, we estimate two OLMs, which can be found in Table 1. Consistent with the SHH, we observe a significant positive interaction effect of the PSA score with self-reported intuitiveness in Model 1 ($\beta = 6.049, p = .021$). That is, an increase in (self-reported) intuitiveness has a greater impact on the probability of prosocial behavior for participants with stronger prosocial attitudes than for participants with less pronounced prosocial attitudes. In addition, we find that intuitiveness does not promote prosocial behavior for participants with low prosocial attitudes ($\beta = -2.943, p = .095$). Furthermore, we find that among perfectly reflective participants variations in prosocial attitudes do not affect prosocial behavior significantly ($\beta = -1.383, p = .399$). Similarly, Model 2 shows the interaction between the PSA score and general intuitiveness; there is a highly significant and positive interaction effect ($\beta = 5.740, p = .005$). Again, we find that general intuitiveness does decrease prosocial behavior for participants without a strong attitude towards prosociality ($\beta = -4.709, p < .001$). It is also observed that among participants with an extreme tendency towards cognitive reflection variations in prosocial attitudes have no impact on prosocial behavior ($\beta = -.329, p = .753$).

Panel (a) and (b) of Fig 4 show the predicted probabilities of behaving prosocially according to Model 1 and Model 2, respectively. For convenience, only the probabilities of the prosocial

Table 1. Ordered logit regression models of behavioral responses in the field experiment.

Response	Model 1		Model 2		Model 3	
	Coef.	SE	Coef.	SE	Coef.	SE
PSA score	-1.383	1.639	-.329	1.045	-3.388 ⁺	1.956
Intuitiveness (self-rep.)	-2.943 ⁺	1.761			-2.541	1.858
PSA × Int. (self-rep.)	6.049 ⁺	2.631			5.626 ⁺	2.749
Intuitiveness (general)			-4.709 ^{***}	1.391	-4.983 ^{**}	1.857
3 PSA × Int. (general)			5.740 ^{**}	2.037	6.171 ⁺	2.741
Disclaimer treatment					.350 ⁺	.191
Theft treatment					.165	.189
High temptation					.152	.189
Male gender					-.155	.210
Age					.342 [*]	.196
Age ²					-.006 ⁺	.004
Naive					.029	.204
McFadden's pseudo R ²	.030		.022		.061	
N	482		744		482	

⁺ $p < .1$,
^{*} $p < .05$,
^{**} $p < .01$,
^{***} $p < .001$.

<https://doi.org/10.1371/journal.pone.0262476.t001>

action and only individuals with an intermediate (= 0.5) or high (= 1) PSA score are shown. Both models predict that participants with high prosocial attitudes behave more prosocially the more intuitive they are, whereas participants with an intermediate PSA score are either not affected by their intuitiveness or behave even less prosocially the greater their tendency towards intuition.

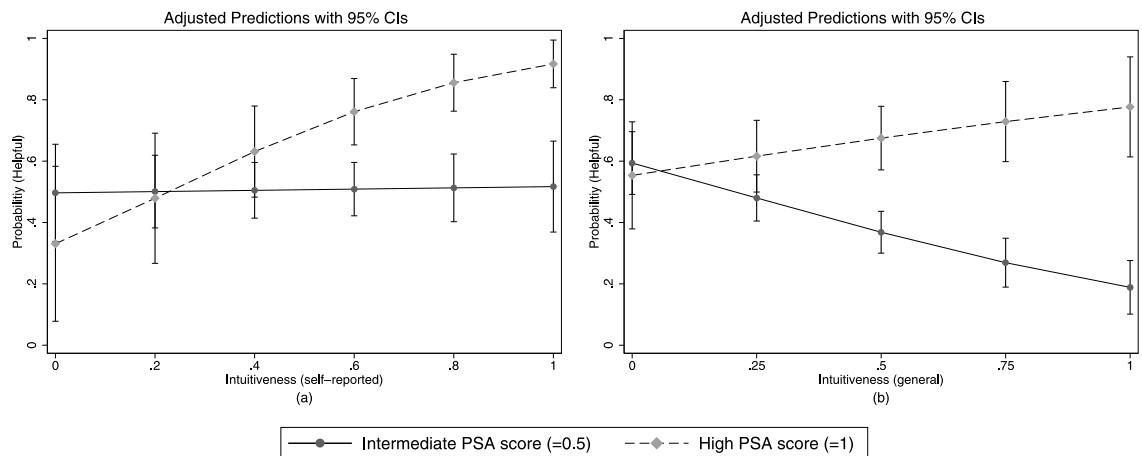


Fig 4. Multivariate analyses. Predicted probabilities of the prosocial action as a function of PSA score and either (a) self-reported intuitiveness or (b) general intuitiveness.

<https://doi.org/10.1371/journal.pone.0262476.g004>

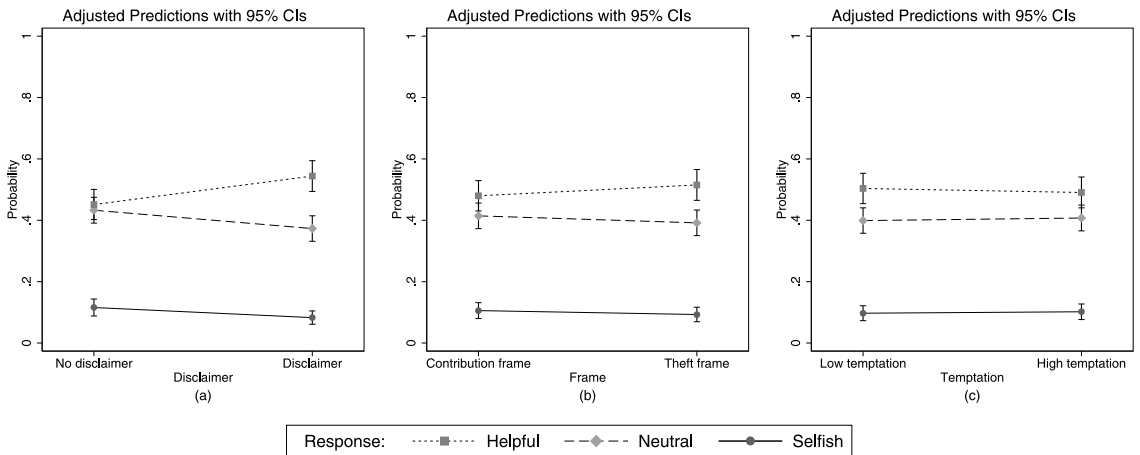


Fig 5. Experimental manipulations. Predicted probabilities of behavioral responses as a function of experimental manipulations.

<https://doi.org/10.1371/journal.pone.0262476.g005>

Let us now turn to the experimental manipulations regarding the content of the email. As before, we use bivariate OLMs for the analyses. We observe that the disclaimer makes it significantly more likely that a more prosocial action is chosen ($\beta = .372, p = .009$, Fig 5(b)). We further observe, that neither the framing of how the imaginary person earned their money ($\beta = .141, p = .317$, Fig 5(b)) nor the amount of money ($\beta = -.052, p = .712$, Fig 5(c)) had any significant effect on the behavior of the participants.

Finally, we combine Model 1 and 2 to perform an additional test for robustness and to check whether both intuitiveness measures have independent effects (Table 1). For this purpose, Model 3 includes the variables of the previous models and additionally controls for our treatment variables (treatment not used = 0 / used = 1). Also, we control for gender (female = 0 / male = 1), previous lab experience (no = 0 / yes = 1), age in years, and squared age.

Our main results of the previous analyses can be replicated in this extended model. Most importantly, we find significant interaction effects between both measures of intuitiveness and the PSA score. Again, neither self-reported nor general intuitiveness promote prosocial behavior among participants with low prosocial attitudes. On the contrary, we even find a significant negative effect of general intuitiveness on prosocial behavior among subjects with an intermediate PSA score. Regarding the experimental manipulations of the content of the email, we observe that the disclaimer treatment has a weakly significant and positive effect on the prosocial behavior of our participants, whereas the theft treatment and the high temptation treatment exhibit no significant effects.

In addition, Model 3 shows that prosocial attitudes have a weakly significant negative effect on prosocial behavior among participants which are extremely reflective according to both measures of intuitiveness. However, this result should not be overestimated, because it is not robust under alternative model specifications.

With respect to the control variables, we observe no significant effect of gender or previous lab experience and only a weakly significant positive effect of age and a weakly significant negative effect of squared age.

Robustness checks

At this point, we perform various robustness checks, to assess whether our results are also valid under other assumptions. First, since the adequacy of the CRT as a measure of intuitiveness has been criticized [45], we also use the number of intuitively answered questions of the extended CRT as an alternative measure for general intuitiveness [12]. We recalculate all models with this alternative index of general intuitiveness and find that all main effects remain virtually unchanged except for the PSA score in the full model which turns insignificant (S5 Table).

Second, since our research interest is focused on interaction effects and these could potentially be biased in the context of logit models [46], we follow standard practice and additionally test our models using linear regression [47]. We observe that the linear models are in line with our previous findings (S6 Table).

Finally, we deal with the problem that is central to many non-reactive studies: How can we ensure that subjects did not suspect they might be part of a study? As mentioned at the beginning of this section, we excluded all subjects from the study who had already expressed suspicion during the field phase. In addition, we asked the participants of the follow-up survey whether they had any suspicions before or during their reaction to the email. A total of 82 subjects indicated that they had developed suspicions before or during their reaction. However, because these subjects do not behave significantly different from those not expressing suspicion in the follow-up survey ($\chi^2 = 3.702, p = 0.157$), we did not remove them from the main analysis. But even if we remove these subjects from the study and recalculate our models, we find no major changes (S7 Table). Altogether, we can say that our results are robust to a number of alternative models.

Discussion

In this paper, we report results of a non-reactive field experiment to contribute to the discussion regarding the social heuristic hypothesis (SHH) [13]. A notable feature of the current study is that the participants of the field experiment had previously participated in a lab experiment which is why we have solid measures of participants' prosocial attitudes as well as their general intuitiveness at our disposal. In addition, we work with two measures of intuitiveness, one based on self-reports in a follow-up survey and the other one based on an extension of the Cognitive Reflection Test [36].

In line with our expectations and previous studies [29, 30, 32], we observe that participants with higher prosocial attitudes behave more prosocial. While we find that self-reported intuitiveness has a positive effect on prosocial behavior in a bivariate analysis, this effect vanishes as soon as we control for prosocial attitudes. In accordance with other studies [12], we find that general intuitiveness has a negative effect on prosocial behavior even in bivariate analysis. What we do find with respect to both self-reported as well as general intuitiveness in multivariate analyses is that prosocial attitudes have a greater impact on prosocial behavior among participants with a greater tendency towards intuitiveness, which is in line with the literature on the SHH and prosocial attitudes [29, 30]. This finding is robust regardless of whether we look at each measure of intuitiveness separately or simultaneously and regardless of whether we work with control variables or not.

All in all, our study contributes to the growing literature that puts serious doubts on the empirical validity of the simplistic hypothesis of intuitive prosociality. As we saw, individual heterogeneity (in our case with respect to internalized attitudes) might be an important factor to consider when studying the interplay of intuitiveness and prosociality. In particular, our results confirm the more nuanced and complex implication of the SHH according to which the

influence of intuitiveness on prosociality is moderated by strength and type of internalized attitudes. Furthermore, we were able to show that the SHH also applies beyond the controlled environment of laboratory studies to a more realistic situation of everyday life.

Regarding the experimental manipulations, on the one hand, we observe that a subtle cue like a disclaimer can influence subjects' responses. On the other hand, the description of how Ms. Koch got her money seems too subtle to cause a behavioral change. In addition, we observe that the incentive associated with the use of the code does not affect prosocial behavior. Taken on face value, this finding contradicts the idea that behavior should be responsive to incentives. However, the finding can also be interpreted differently against the background of experimental literature on lying and cheating [48, 49]. Regarding the act of deception, [50] shows that while increasing the gain for the actor increases the probability of deception, increasing the harm to another person decreases this probability. Since in our field experiment both potential gain and harm were de facto manipulated simultaneously, the effects of these two kinds of incentives could neutralize each other, which might explain why we observe no effect.

Turning to the limitations of the current study, we first discuss a peculiar drawback of our non-reactive design. With respect to participants who showed a neutral reaction, we have no way of knowing whether they actually got the email or not. As a consequence, the fraction of neutral reactions to our email is potentially biased upwards. While we acknowledge this drawback, we do not believe that it seriously diminishes our qualitative findings, for several reasons. First, note that only our analyses working with the general measure of intuitiveness are affected by this potential problem. The results with respect to the self-reported measure only refer to subjects who participated in the follow-up survey and were consequently able to get the email from our lab. Also, unreported, additional analyses show that our results regarding general intuitiveness are robust if we restrict the analyses to those subjects who participated in the follow-up survey. Second, there are good reasons to believe that the fraction of participants who did not react to the email because they have not read it, is actually rather small. For one, almost all of our participants were students who subscribed to the email list of the lab voluntarily, most likely to earn some extra cash. Against this background, it is reasonable to expect our participants to be rather attentive towards emails from the lab. Also, the participation rates regarding the follow-up survey do not differ too much between participants who showed no reaction and those who showed a selfish or helpful reaction.

Other limitations of the current study are related to the special composition of the sample and the measurements of intuitiveness. First, there are almost no participants with low prosocial attitudes. As a consequence, we could not test the interesting hypothesis that low prosocial attitudes have a greater impact on selfish behavior among participants with a stronger tendency towards intuitiveness. Future research should work with a more diverse sample to remedy this deficit. Finally, it was a bit surprising to discover that our two measures of intuitiveness do not correlate at all. While it can even be interpreted as an advantage, because our results are robust with respect to two independent measures of intuitiveness, it raises the question to what extent the measure based on self-reports is actually reliable. In any case, it is a worthwhile task for future research to check whether our findings regarding the interaction between prosocial attitudes and intuitiveness are robust if intuitiveness is not measured but experimentally manipulated, for instance via time pressure or cognitive overload.

Supporting information

S1 Data.
(CSV)

S1 Table. Email text. The text of the email translated from German. Text passages in square brackets mark experimental manipulations.

(PDF)

S2 Table. Self-reported intuitiveness. German translations and scales of questions to determine the intuitiveness of the decision in the field experiment.

(PDF)

S3 Table. Short Prosocial Personality Battery. All responses to these questions were answered on a 5-point Likert scale ranging from 1 “Strongly Disagree” to 5 “Strongly Agree”.

(PDF)

S4 Table. Extended cognitive reflection test.

(PDF)

S5 Table. Alternative ordered logit regression models. Using the alternative measure for general intuitiveness.

(PDF)

S6 Table. Linear regression models. The response is treated as quasi metric ranging from -1 (selfish) to 0 (neutral) to 1 (helpful).

(PDF)

S7 Table. Ordered logit regression models. Without subjects that stated suspicion in the follow-up survey.

(PDF)

Acknowledgments

The authors thank Anna-Luise Schönheit and Daniel Peter for proofreading earlier drafts of the manuscript as well as two anonymous reviewers for their valuable input.

Author Contributions

Conceptualization: Sascha Grehl, Andreas Tutić.

Data curation: Sascha Grehl, Andreas Tutić.

Formal analysis: Sascha Grehl, Andreas Tutić.

Funding acquisition: Sascha Grehl, Andreas Tutić.

Investigation: Sascha Grehl, Andreas Tutić.

Methodology: Sascha Grehl, Andreas Tutić.

Project administration: Sascha Grehl, Andreas Tutić.

Visualization: Sascha Grehl.

Writing – original draft: Sascha Grehl, Andreas Tutić.

References

1. Henrich J, Boyd R, Bowles S, Camerer CF, Fehr E, Gintis H, et al. In Search of Homo Economicus: Behavioral Experiments in 15 Small-Scale Societies. *American Economic Review*. 2001; 91(2):73–78. <https://doi.org/10.1257/aer.91.2.73>
2. Fehr E, Gächter S. Altruistic Punishment in Humans. *Nature*. 2002; 415(6868):137–140. <https://doi.org/10.1038/415137a> PMID: 11805825

3. Camerer CF. Behavioral Game Theory: Experiments in Strategic Interaction. Princeton, NJ: Princeton University Press; 2003.
4. Kahneman D. Thinking, Fast and Slow. London, UK: Penguin Books; 2011.
5. Evans J. Thinking Twice: Two Minds in One Brain. Oxford, UK: Oxford University Press; 2010.
6. Stanovich KE. Rationality and the Reflective Mind. New York, NY: Oxford University Press; 2011.
7. Evans JSBT, Stanovich KE. Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives on Psychological Science*. 2013; 8(3):223–241. <https://doi.org/10.1177/1745691612460685> PMID: 26172965
8. Rand DG, Greene JD, Nowak MA. Spontaneous Giving and Calculated Greed. *Nature*. 2012; 489(7416):427–430. <https://doi.org/10.1038/nature11467> PMID: 22996558
9. Zaki J, Mitchell JP. Intuitive Prosociality. *Current Directions in Psychological Science*. 2013; 22(6):466–470. <https://doi.org/10.1177/0963721413492764>
10. Rubinstein A. Instinctive and Cognitive Reasoning: A Study of Response Times. *Economic Journal*. 2007; 117(523):1243–1259. <https://doi.org/10.1111/j.1468-0297.2007.02081.x>
11. Nielsen UH, Tyrann JR, Wengström E. Second Thoughts on Free Riding. *Economics Letters*. 2014; 122(2):136–139. <https://doi.org/10.1016/j.econlet.2013.11.021>
12. Peysakhovich A, Rand DG. Habits of Virtue: Creating Norms of Cooperation and Defection in the Laboratory. *Management Science*. 2016; 62(3):631–647. <https://doi.org/10.1287/mnsc.2015.2168>
13. Rand DG, Peysakhovich A, Kraft-Todd GT, Newman GE, Wurzbacher O, Nowak MA, et al. Social Heuristics Shape Intuitive Cooperation. *Nature Communications*. 2014; 5(1):3677. <https://doi.org/10.1038/ncomms4677> PMID: 24751464
14. Rand DG, Kraft-Todd GT. Reflection Does Not Undermine Self-Interested Prosociality. *Frontiers in Behavioral Neuroscience*. 2014; 8:300. <https://doi.org/10.3389/fnbeh.2014.00300> PMID: 25232309
15. Gilbert DT, Pelham BW, Krull DS. On Cognitive Busyness: When Person Perceivers Meet Persons Perceived. *Journal of Personality and Social Psychology*. 1988; 54(5):733–740. <https://doi.org/10.1037/0022-3514.54.5.733>
16. Schulz JF, Fischbacher U, Thöni C, Utikal V. Affect and Fairness: Dictator Games Under Cognitive Load. *Journal of Economic Psychology*. 2014; 41:77–87. <https://doi.org/10.1016/j.joep.2012.08.007>
17. Baumeister RF, Bratslavsky E, Muraven M, Tice DM. Ego Depletion: Is the Active Self a Limited Resource? *Journal of Personality and Social Psychology*. 1998; 74(5):1252–1265.
18. Krajbich I, Bartling B, Hare T, Fehr E. Rethinking Fast and Slow Based on a Critique of Reaction-Time Reverse Inference. *Nature Communications*. 2015; 6:7455. <https://doi.org/10.1038/ncomms8455> PMID: 26135809
19. Evans AM, Dillon KD, Rand DG. Fast but Not Intuitive, Slow but Not Reflective: Decision Conflict Drives Reaction Times in Social Dilemmas. *Journal of Experimental Psychology: General*. 2015; 144(5):951–966. <https://doi.org/10.1037/xge0000107>
20. Corgnet B, Espín AM, Hernán-González R. The Cognitive Basis of Social Behavior: Cognitive Reflection Overrides Antisocial But Not Always Prosocial Motives. *Frontiers in Behavioral Neuroscience*. 2015; 9:287. <https://doi.org/10.3389/fnbeh.2015.00287> PMID: 26594158
21. Capraro V, Corgnet B, Espín AM, Hernán-González R. Deliberation Favours Social Efficiency by Making People Disregard Their Relative Shares: Evidence From USA and India. *Royal Society Open Science*. 2017; 4(2):160605. <https://doi.org/10.1098/rsos.160605> PMID: 28386421
22. Tinghög G, Andersson D, Bonn C, Böttiger H, Josephson C, Lundgren G, et al. Intuition and Cooperation Reconsidered. *Nature*. 2013; 498(7452):E1–E2. <https://doi.org/10.1038/nature12194> PMID: 23739429
23. Verhoeven PPJL, Bouwmeester S. Does Intuition Cause Cooperation? *PLOS ONE*. 2014; 9:e96654. <https://doi.org/10.1371/journal.pone.0096654> PMID: 24801381
24. Bouwmeester S, Verhoeven PPJL, Aczel B, et al. Registered Replication Report: Rand, Greene, and Nowak (2012). *Perspectives on Psychological Science*. 2017; 12(3):527–542. <https://doi.org/10.1177/1745691617693624> PMID: 28475467
25. Bear A, Rand DG. Intuition, Deliberation, and the Evolution of Cooperation. *PNAS*. 2016; 113(4):936–941. <https://doi.org/10.1073/pnas.1517780113> PMID: 26755603
26. Tomlin D, Rand DG, Ludvig EA, Cohen JD. The Evolution and Devolution of Cognitive Control: The Costs of Deliberation in a Competitive World. *Scientific Reports*. 2015; 5:11002. <https://doi.org/10.1038/srep11002> PMID: 26078086
27. Toupo DFP, Strogatz SH, Cohen JD, Rand DG. Evolutionary Game Dynamics of Controlled and Automatic Decision-Making. *Chaos*. 2015; 25:073120. <https://doi.org/10.1063/1.4927488> PMID: 26232971

28. Rand DG, Brescoll VL, Everett JAC, Capraro V, Barcelo H. Social Heuristics and Social Roles: Intuition Favors Altruism for Women but not for Men. *Journal of Experimental Psychology: General*. 2016; 145(4):389. <https://doi.org/10.1037/xge0000154>
29. Mischkowski D, Glöckner A. Spontaneous Cooperation for Prosocials, But Not for Proselfs: Social Value Orientation Moderates Spontaneous Cooperation Behavior. *Scientific Reports*. 2016; 6(1):1–5. <https://doi.org/10.1038/srep21555>
30. Yamagishi T, Matsumoto Y, Kiyonari T, Takagishi H, Li Y, Kanai R, et al. Response Time in Economic Games Reflects Different Types of Decision Conflict for Prosocial and Proself Individuals. *Proceedings of the National Academy of Sciences*. 2017; 114(24):6394–6399. <https://doi.org/10.1073/pnas.1608877114> PMID: 28559334
31. Santa JC, Exadaktylos F, Soto-Faraco S. Beliefs About Others' Intentions Determine Whether Cooperation is the Faster Choice. *Scientific Reports*. 2018; 8(1):1–10.
32. Andrighetto G, Capraro V, Guido A, Szekeley A. Cooperation, Response Time, and Social Value Orientation: A Meta-Analysis. Working Paper;. Available from: <https://psyarxiv.com/cbakz>.
33. Alós-Ferrer C, Garagnani M. The Cognitive Foundations of Cooperation. *Journal of Economic Behavior and Organization*. 2020; 175:71–85. <https://doi.org/10.1016/j.jebo.2020.04.019>
34. Author A, Author B, Author C, Author D. A Sociological Dual-Process Action Theory: Experimental Evidence on Prosocial Behavior. Working paper; 2021.
35. Batson CD, Powell AA. Altruism and Prosocial Behavior. In: Millon T, Weiner IB, editors. *Handbook of Psychology*. Hoboken, NJ: Wiley; 2003. p. 463–484.
36. Frederick S. Cognitive Reflection and Decision Making. *Journal of Economic Perspectives*. 2005; 19(4):25–42. <https://doi.org/10.1257/089533005775196732>
37. Milgram S, Mann L, Harter S. The Lost-Letter Technique: A Tool of Social Research. *Public Opinion Quarterly*. 1965; 29(3):437. <https://doi.org/10.1086/267344>
38. Farrington DP, Knight BJ. Two Non-Reactive Field Experiments on Stealing from a 'Lost' Letter. *British Journal of Social and Clinical Psychology*. 1979; 18(3):277–284. <https://doi.org/10.1111/j.2044-8260.1979.tb00337.x> PMID: 519138
39. Penner LA, Fritzsche BA, Craiger P, Freifeld TR. Measuring the Prosocial Personality. In: Butcher JN, Spielberg CD, editors. *Advances in Personality Assessment*. vol. 10. New York, NY: Psychology Press; 1995. p. 147–163.
40. Brañas-Garza P, Kujal P, Lenkei B. Cognitive Reflection Test: Whom, How, When. *Journal of Behavioral and Experimental Economics*. 2019; 82:101455. <https://doi.org/10.1016/j.socec.2019.101455>
41. Thomson KS, Oppenheimer DM. Investigating an Alternate Form of the Cognitive Reflection Test. *Judgment and Decision Making*. 2016; 11(1):99.
42. Toplak ME, West RF, Stanovich KE. The Cognitive Reflection Test as a Predictor of Performance on Heuristics-and-Biases Tasks. *Memory & Cognition*. 2011; 39(7):1275–1289. <https://doi.org/10.3758/s13421-011-0104-1> PMID: 21541821
43. Williams R. Understanding and Interpreting Generalized Ordered Logit Models. *Journal of Mathematical Sociology*. 2016; 40(1):7–20. <https://doi.org/10.1080/0022250X.2015.1112384>
44. Brant R. Assessing Proportionality in the Proportional Odds Model for Ordinal Logistic Regression. *Biometrics*. 1990; 46(4):1171–1178. <https://doi.org/10.2307/2532457> PMID: 2085632
45. Pennycook G, Cheyne JA, Koehler DJ, Fugelsang JA. Is the Cognitive Reflection Test a Measure of Both Reflection and Intuition? *Behavior Research Methods*. 2016; 48(1):341–348.
46. Ai C, Norton EC. Interaction Terms in Logit and Probit Models. *Economics Letters*. 2003; 80(1):123–129. [https://doi.org/10.1016/S0165-1765\(03\)00032-6](https://doi.org/10.1016/S0165-1765(03)00032-6)
47. Breen R, Karlson KB, Holm A. Interpreting and Understanding Logits, Probits, and Other Nonlinear Probability Models. *Annual Review of Sociology*. 2018; 44:39–54. <https://doi.org/10.1146/annurev-soc-073117-041429>
48. Mazar N, Amir O, Ariely D. The Dishonesty of Honest People: A Theory of Self-Concept Maintenance. *Journal of Marketing Research*. 2008; 45(6):633–644. <https://doi.org/10.1509/jmkr.45.6.633>
49. Fischbacher U, Föllmi-Heusi F. Lies in Disguise—An Experimental Study on Cheating. *Journal of the European Economic Association*. 2013; 11(3):525–547. <https://doi.org/10.1111/jeea.12014>
50. Gneezy U. Deception: The Role of Consequences. *American Economic Review*. 2005; 95(1):384–394. <https://doi.org/10.1257/0002828053828662>

**Implizite Einstellungen, explizite Einstellungen
und die Affinität zur AfD (Implicit Attitudes,
Explicit Attitudes, and the Affinity Towards the
AfD)**

Tutić, Andreas and Sascha Grehl

Kölner Zeitschrift für Soziologie und Sozialpsychologie 73: 389-417 (2021). Translated from German.

Implizite Einstellungen, explizite Einstellungen und die Affinität zur AfD*

Andreas Tutić[†]
Sascha Grehl[‡]

Implicit attitudes, explicit attitudes and the affinity toward the AfD

Abstract

Most electoral studies and especially empirical research regarding the question of how voting intentions for the AfD can be explained take into account the influence of cultural orientations, but only in the form of explicit attitudes, which are measured by evaluative verbal expressions. Against the background of the dual-process perspective, this paper argues that, in addition to explicit attitudes, implicit attitudes, which represent associative links between mentally represented attitude objects and their evaluation, are not to be neglected in explaining voting intentions for the AfD. According to the principle of catalyzation, implicit attitudes are more strongly reflected in explicit attitudes and also in overt behavior when the articulation of explicit attitudes or behavior takes place in an intuitive rather than reflected cognitive process. These action-theoretic ideas are tested in an exploratory study with 960 respondents. We find that both implicit and explicit attitudes toward populism and racism influence the intention to vote for the AfD. Also, as predicted by the principle of catalyzation, the influence of implicit attitudes depends on whether respondents are more inclined to intuitive or reflective cognitive processes.

Keywords: attitudes, dual process perspective, intuition, AfD, voting behavior

1 Introduction

Recent studies indicate that the prevalence of racist and xenophobic attitudes in Germany has been declining in recent years (Zieck et al. 2019; Decker et al. 2020). Based on the “Mitte” study by the Friedrich Ebert Foundation, there has been a discernible decrease in the prevalence of racist and xenophobic attitudes in Germany over the past few years. Specifically, in 2002, 12.2% of respondents agreed with racist statements, and 34.5% with xenophobic statements, while in 2019, the corresponding figures were found to be 7.2% and 18.8%, respectively (Zieck et al. 2019, p. 82). As pleasing as this development may appear at first glance, the finding is questionable in view of the recent rise of serious right-wing extremist-motivated acts of terrorism and violence, such as those in Hanau or Halle. The impression of an increase in politically motivated criminal and violent acts with a right-wing extremist background is also confirmed by the official statistics from the Federal Criminal Police Office. While in 2009, the first year of the survey, there were 18750 such acts in Germany according to the official count, the statistics for 2019 show the number 21290 (Statista 2020).

Now it could be argued that it is quite possible that a decrease in right-wing extremist attitudes in the population is accompanied by an increase in right-wing extremist crimes. Using Homans' (1974) frustration-aggression hypothesis, it could be argued, for example, that isolated right-wing extremists act more aggressively the more their “convictions” lose support among the population. Another recent development in the Federal Republic of Germany, however, makes the claim of the decline of racist and xenophobic attitudes seem quite implausible: The rise of the AfD, which since 2017 has been the (numerically) strongest faction of the opposition in the Bundestag and in some federal states is competing for the status of the strongest party; a party that, according to the sociological meaning of the term, is right-wing populist in every sense of the word (Mudde 2007) and may soon be classified by the Federal Office for the Protection of the Constitution as a suspected right-wing extremist party.

How is it possible that, on the one hand, racist and xenophobic attitudes are losing ground in the broad population, while at the same time a right-wing populist party that elevates precisely these attitudes to a kind of political program is taking root in the political system of the republic? This paper aims to help clarify this question. To do so, we draw on the dual-process perspective (Kahneman 2011; Evans 2010; Stanovich 2011; Esser and Kroneberg 2020) to distinguish between two types of attitudes, namely implicit and explicit attitudes (Wilson et al. 2000; Wilson 2002). For example, the “Mitte” study previously cited uses only explicit procedures to measure racist and xenophobic attitudes. In doing so, respondents express their agreement or disagreement for a series of statements, such as “Whites are rightly leading the world.” using a multilevel Likert scale (Zieck et al. 2019, pp. 70ff.). From an action-theoretic perspective, such explicit attitudes are merely artifacts, meaning that while they are based on underlying implicit attitudes, they are additionally influenced by situational and motivational considerations, such as social desirability. This is not just a matter of respondents falsifying their “true” attitudes for strategic motives and deliberately making “false” statements. Rather, implicit and explicit attitudes are anchored in different segments or sectors of the respondents' memory. Respondents do not necessarily need to know about their implicit attitudes and therefore cannot articulate them *expressis verbis* at

all, be it in standardized surveys or even in qualitative in-depth interviews (cf. Vaisey 2014). Implicit attitudes are based on associative links between attitude objects and a positive or negative evaluation and can be made visible by techniques borrowed from social and cognitive psychology, such as the Implicit Association Test (Fazio et al. 1995; Greenwald et al. 1998). In light of this, we argue that the dual-process perspective incorporates a central action-theoretic principle of catalyzation, which states that implicit attitudes should be reflected more strongly in the articulation of explicit attitudes and in overt behavior the more respondents tend to act in an intuition-guided manner.

The relevance of implicit attitudes as well as the principle of catalyzation are validated by means of an explorative study with 960 respondents.¹ It is shown that implicit attitudes toward populism and racism are only rather loosely related to the corresponding explicit attitudes, whereby this relationship is strengthened among respondents who are more inclined toward intuition. Furthermore, both implicit attitudes also influence affinity to the AfD (operationalized via voting intention) - controlling for the standard variables from the relevant research as well as for both explicit attitudes. In this context, too, the principle of catalyzation is confirmed: the more respondents tend to rely on their intuitive gut feeling, the more strongly implicit populism and implicit racism are reflected in the probability of voting for the AfD.

Due to the fact that we only have cross-sectional data, which was furthermore collected with a convenience sample from an access panel provider, we cannot make any firm statements about the prevalence of implicit and explicit attitudes in Germany over time. However, the present study makes clear that implicit and explicit attitudes are only loosely related and that implicit attitudes can have a strong effect on affinity for right-wing populist parties under specifiable conditions. Against the backdrop of these findings, it is entirely possible that perhaps public discourse in Germany has increasingly shed racist and xenophobic attitudes in recent decades, but at the same time, behind or under the veil of consciousness, so to speak, deep-seated implicitly racist attitudes continue to be prevalent or have even increased, enabling the electoral successes of the AfD.

The paper is organized as follows: Section 2 lays the theoretical foundations. Section 3 describes our methodological approach, in particular the instrument for measuring implicit attitudes borrowed from psychology. In Section 4, we present our empirical findings. Section 5 places these findings in the context of the current state of research, discusses any limitations of the study, and suggests possible follow-up projects.

¹ We refer to this as an exploratory study for several reasons: It is a non-preregistered, non-experimental survey of a convenience sample. Furthermore, we use a measure for intuition-guided action that is established in research on socially desirable behavior, but is not a standard measure of intuitiveness.

2 Dual-Process Perspective and the Principle of Catalyzation

We unfold our somewhat involved action-theoretical argument in several steps. First, we introduce the main ideas of the dual-process perspective. We then explain the conceptual distinction between implicit and explicit attitudes and the principle of catalyzation. Finally, we present what can be learned from these considerations in terms of empirically testable hypotheses in the context of this study.

2.1 Dual Process Perspective

The dual-process perspective (DPP) emerges from the substantive convergence of three strands of literature: The recent sociological theories of action on framing and variable rationality (Esser 1996; Kroneberg 2005; Esser and Kroneberg 2020; Lindenberg 2013), the dual-process approach in cognitive and social psychology (Kahneman 2011; Evans 2010; Stanovich 2011), and the axiomatic theories of bounded rationality (Rubinstein 1998, 2013; Tutić 2015a, 2015b). For the purposes of this paper, it is sufficient to outline some of the most important basic ideas of the DPP.

First, the behavior of an actor² is explained from the interaction of two “selves”. Different authors use different terminologies, such as modes (Esser, Kroneberg, Fazio), systems (Kahneman), types of processes (Evans and Stanovich) or reflective versus intuitive decision-making (Rubinstein). Widely shared, however, is the description of the psychological characteristics of these two selves.³ Type 1 processes are usually automatic, fast, based on associations, and occur outside the actor's awareness. Type-2 processes, on the other hand, are controlled, slow, based on calculations, and are processed within the actor's consciousness. Empirical evidence for *qualitative* differences between Type 1 and Type 2 processes can be brought to light through study designs that involve, for example, time pressure or working memory overload to make Type 2 processes difficult or even impossible (e.g., Rand et al. 2012; Rand and Kraft-Todd 2014). Moreover, neuroscientific imaging techniques reveal that Type 1 and Type 2 processes are associated with characteristic differences in neural activity in different areas of the brain (e.g., McClure et al. 2004; Greene et al. 2001).

Second, it depends to a large extent on the definition of the situation whether an action is chosen within a framework of Type 1 or Type 2 processes. The model of frame selection

² For reasons of linguistic economy, we occasionally resort to the generic masculine without wishing to express a gender bias. (This footnote only makes sense in German, as the German language distinguishes between male and female actors.)

³ The notion of two “selves” should not be philosophically exaggerated. From a decision-theoretical perspective, many of the decision processes studied in DPP are in the tradition of “multiple-selves models” (Strotz 1955; Tutić 2015a). In the following, we will generally speak of Type 1 and Type 2 processes and only occasionally and for linguistic reasons resort to alternative formulations, such as intuitive and reflective.

(MFS; Esser 1996; Kroneberg 2005) as a special representative of the DPP expresses this second idea particularly clearly. According to this model, actors possess mental schemata of categorization, so-called frames, which transform objective situations into subjectively perceived situations. Frames exhibit varying degrees of congruence with an objective situation. The adequacy of a frame is determined not only by objectively ascertainable cues, such as the presence of salient objects, but also by the actor's internal dispositions, specifically the chronic or transient accessibility of the frame. If the degree of congruence between a frame and an objective situation, the so-called match, is sufficiently high, the definition of the situation emerges spontaneously, without requiring a conscious and reflective process of interpretation. If the actor additionally has mental schemata of routine action sequences, so-called scripts, and if the fit of a script (the so-called activation weight) to a previously selected frame is sufficiently high, then an automatic-spontaneous selection of this script occurs. Furthermore, if there is an action available that is sufficiently suitable for the previously selected script, then the observable action will also be selected automatically-spontaneously. Thus, whether an action is selected as part of a Type 1 or Type 2 process depends on how well the intuitively accessible frame fits the objective situation.

Third, the DPP assumes that the majority of individual actions and social interactions occur within the framework of automatic Type 1 processes. Conscious and reflective interpretation of situations and deliberate consideration of alternative actions based on expected consequences, as well as the design of action plans, represent a borderline case that only arises when the actor is placed in a sufficiently unusual situation and has the opportunity and motivation to reflect. In this way, the DPP shares a central moment of Alfred Schütz's phenomenological theory and also of practice theory, for instance in the version of Bourdieu (1990) and Giddens (1984), who emphasize the empirical relevance of the lifeworld as well as the *doxa* and oppose intellectualization in the explanation of behavior and action (cf. Reckwitz 2003).⁴

Finally, according to the DPP, cultural orientations, in particular attitudes, values and internalized norms, exert their influence on behavior through distinguishable cognitive mechanisms. In the case of the MFS, for example, it can be argued that on the one hand they influence the definition of the situation (frame, match, accessibility), and, on the other hand, they have an affinity with forms of routine action sequences (scripts). Crucial for the motivational power and meaning-giving potential of cultural orientations is that they can influence action within the framework of Type-1 processes through the selection of frames and scripts. In addition, the MFS also allows cultural orientations to have an effect on action in the context of Type 2 processes. For instance, actors can reason about the situational validity and adequacy of cultural orientations in the context of reflective interpretation processes (Kroneberg 2007). Furthermore, conformity to cultural orientations can be one of the consequences that the actor takes into account when making a reflective decision about concrete actions. The thesis that cultural orientations are reflected in behavior through different cognitive mechanisms has proved exceedingly fruitful in American cultural sociology (cf.

⁴ The parallels between the DPP and practice-theoretical approaches are by no means exhausted in this point and require systematic elaboration.

DiMaggio 1997; Martin 2010). Vaisey (2009), for example, argues that two classical views of the role of cultural orientations can be brought into some kind of synthesis with the DPP. First, the older sociological position of Weber ([1921/1922] 2002), Durkheim ([1893] 1997), and Parsons (1937), according to which cultural orientations have a direct impact as a motive for action. And second, the idea, formulated in particular by Swidler (1986), that cultural orientations are rather a kind of toolbox from which actors make strategic use, also in order to rationalize and legitimize actions ex-post in discourse, which in doubt may be motivated in a completely different way. With regard to the DPP, it can be said that cultural orientations function as a direct motive for action when they exert their influence in the context of Type 1 processes. Strategic use of cultural orientations and discursive rationalizations with the help of cultural orientations refer more to Type 2 processes.

In our study, we work with a dispositive measure of respondents' tendency to rely on intuitive Type 1 processes (see Section 3), which we also call the measure of intuition-guided action. In theoretical terms, respondents who score high on this measure can be conceptualized as actors who are chronically exposed to high reflection costs and are therefore more likely to act intuitively.

2.2 Implicit Attitudes, Explicit Attitudes and the Principle of Catalyzation

In this paper, we will take a closer look at a specific type of cultural orientation, namely attitudes. An attitude toward an attitude object can be defined as a psychological tendency to evaluate that object positively or negatively (Eagly and Chaiken 1993; Fazio 1995). Attitudes can be measured via implicit or via explicit measures, in this case we referred to them either as implicit and explicit attitudes (Wilson et al. 2000; Fazio and Olson 2003; Strack and Deutsch 2004; Rydell and McConnell 2006). Implicit measures rely on making mentally anchored associative links, characteristic of System 1 (Kahneman 2011), between an attitude object and an evaluation visible in the context of Type 1 processes. The Brief Implicit Association Test used in this study (Sriram and Greenwald 2009; Nosek et al. 2014), for instance, focuses on the comparison of response latencies in matching tasks in which objects with positive or negative attributes are matched (see Section 3). Explicit measures for attitudes operate with explicit, (latent) evaluative expressions. Respondents express their agreement or disagreement with evaluative statements in the context of a standardized survey or make evaluative statements on their own initiative in the context of qualitative interviews.

While implicit attitudes by definition become visible in the context of Type 1 processes, explicit attitudes can be explicated in the context of both Type 1 and Type 2 processes. When explicit attitudes are expressed in the context of a Type 2 process, additional considerations, such as an interest in articulating socially desirable attitudes, may enter into the explication in addition to the underlying implicit attitudes (Fazio et al. 1995). Explicit attitudes are thus situationally and motivationally biased representations of implicit attitudes. Implicit and explicit attitudes toward the same attitude object are therefore generally not particularly closely related (Wilson et al. 2000). Respondents can express themselves consistently *expressis verbis* positively (or negatively) toward an attitude object and at the same time have incorporated a negative (or positive) implicit attitude toward the attitude object. Another interesting aspect of the distinction between implicit and explicit attitudes is that

actors are generally much more aware that they are expressing an explicit attitude than they are aware that they are the bearers of an implicit attitude (Rydell and McConnell 2006).

In this context, the DPP allows for the formulation of a general action-theoretical principle. According to the principle of catalyzation, implicit attitudes are more strongly reflected in explicit attitudes as well as in observable behavior when this articulation of explicit attitudes or behavior takes place in the context of Type 1 processes. This is because implicit attitudes have the potential for this case to dominate the definition of the situation and the selections of script and action as a guiding principle, whereas their influence is diluted by situational considerations in the context of reflective Type 2 processes. Conformity to an implicit attitude is, in the reflective-calculative mode, only one among several consequences of action that the actor considers.⁵

The influence of explicit attitudes on action is much more difficult to theorize. On the one hand, an explicit attitude is a proxy for an underlying implicit attitude and should therefore have a stronger influence on behavior when that behavior occurs as part of a Type 1 process. On the other hand, explicit attitudes may also deviate from underlying implicit attitudes, especially when the articulation of explicit attitudes occurs in the context of a Type 2 process. Thus, when there is a strong divergence, their influence on behavior that occurs automatically-spontaneously should diminish. Explicit attitudes are inherently artifacts from the DPP perspective, which poses major challenges to theory building and empirical research. In this paper, we therefore focus on the explanatory potential of implicit attitudes.

2.3 Empirically Testable Hypotheses

As already made clear in the introduction, in this paper we are interested in how implicit attitudes, translate into explicit attitudes and ultimately into voting intentions for the AfD, considering actors' cognitive disposition toward intuitiveness as a central moderator variable. Following existing research on right-wing populist tendencies (e.g., Kitschelt 1995; Ivarsson 2005; Werts et al. 2013) and the German literature on the AfD (Lengfeld 2017; Lux 2018; Tutić and Hermann 2018; Rippl and Seipel 2018; Lengfeld and Dilger 2018), we focus on attitudes toward racism and populism.

Starting from the principle of catalyzation, one arrives at empirically testable hypotheses when incorporating theoretical ideas about the conditions for the onset of Type 1 or Type 2 processes. In this context, the notion of so-called default-interventionist model dominates in the literature (cf. Evans and Stanovich 2013). According to this model, system 1 automatically offers a definition of the situation and also impulses for action (default), which are usually accepted by system 2 without question. In exceptional cases, system 2 intervenes and, if

⁵ Empirical research on MFS usually focuses on a second fundamental principle of action theory (e.g., Kroneberg et al. 2010). According to the *principle of suppression*, (anticipated) pure consequences of action, i.e., aspects of decision situations that are not anchored in the "associative machinery" (Kahneman 2011) of system 1, should exert a stronger influence on action when this action occurs in the context of Type 2 processes.

cognitive capacities such as working memory are available, overwrites the impressions and impulses generated by system 1 with the results of its own processing (override). Mode selection in the context of MFS expresses the conditions for this intervention very clearly in comparison to alternative conceptions: the greater the match between mental model and objective situation, the lower the chances of success and the potential yield of reflection, and the higher the costs of reflection, the less likely it is that a Type 2 process will occur (Kroneberg 2005, 2011).

Specifically, in this study we are interested in the following hypothesis:

Among respondents who tend to be more intuitive, implicit attitudes toward populism and racism are more strongly reflected in explicit attitudes toward populism and racism and in affinity for the AfD than among respondents who tend to be more reflective.

We would like to point out that this hypothesis implies two ideas that are not necessarily self-evident tenets of empirical social research and should therefore be emphasized here. First, implicit and explicit attitudes can diverge considerably under certain conditions, which is why explicit attitudes are only of limited use as proxies for implicit attitudes. Second, under certain conditions, implicit attitudes are significant predictors of sociologically relevant phenomena. Both statements and the stated hypothesis are supported by the empirical study described below.

3 Data, Instruments, and Variables

Based on a convenience sample from the access panel operator respondi, 1102 people were surveyed in June 2020. The survey was conducted online and took approximately 30 minutes to complete. We used a population-representative approach to quota the survey participants based on their age and gender. However, we intentionally oversampled individuals residing in Eastern Germany, comprising approximately half of our sample. This was done against the background that the electoral base for the AfD is particularly strong in the east and we wanted to ensure that enough AfD voters were available for later analysis.

Brief Implicit Association Test

First, we would like to introduce the centerpiece of our study, the so-called Brief Implicit Association Test (BIAT), in more detail. The BIAT (Sriram and Greenwald 2009; Nosek et al. 2014) is a condensed version of the classic Implicit Association Test (Greenwald et al. 1998; Fazio et al. 1995). A total of four word lists are required for administration. One of these word lists consists exclusively of words with positive (“good”) and one exclusively of words with negative (“bad”) connotations. In addition, two lists are required, each of which contains words associated with the attitude objects to be compared. In this study, we administered two BIATs that differ in the attitude objects used. The BIAT measuring implicit racism uses

the attitude objects “Refugees” and “Germans”.⁶ The BIAT for measuring implicit populism uses the attitude objects “Established” and “Alternative”.

A BIAT consists of two blocks, which differ in the attitude object that is focal. While the attribute “good” is focal in both blocks, the attribute “bad” is never focal. In each block, respondents are first presented with the focal attribute, the focal attitude object, and the word lists associated with each. In the actual test, participants are presented with each word from the four word lists in random order. The words associated with the attribute and the attitude object alternate constantly. For each word, participants must quickly decide whether it belongs to the focal attribute or attitude object, or not.⁷ Figure 1 shows such a matching task: the focal attitude object in this case is “Germans” and the focal attribute is “good”. The respondent must decide whether the presented word (“Thomas”) is associated with “Germans” or “good” (by pressing “K”) or with neither (by pressing “D”).

The idea behind this approach is that respondents who have a positive implicit attitude toward the focal attitude object should take less time to assign than respondents who have a negative implicit attitude toward the focal attitude object.⁸ This is because the attribute “good” is always focal by design. According to Greenwald et al. (2003), the different reaction times resulting from the two blocks can be used to calculate the so-called *D*-measure, which has been shown to be superior to alternative measures (cf. Sriram et al. 2006).⁹ The *D*-measure is the difference between the mean latencies of the two BIAT blocks divided by the standard deviation of the latencies in the two blocks.¹⁰ Typically, these values are between

⁶ Strictly speaking, this is not a measure of implicit racism, but a measure of implicit negative attitudes toward refugees. However, we assume that these negative attitudes toward refugees are based on implicit racist and xenophobic attitudes.

⁷ In case of an error, i.e., a wrong classification, a red X appears and the word is presented until it has been correctly classified.

⁸ The use of latency in the context of the DPP has a long tradition in German sociology (Stocké 2004; Urban and Mayerl 2007; Beier 2016). Here, however, the response times that respondents need to express their explicit attitudes are typically recorded.

⁹ In recent years, a number of critical contributions to the interpretation of reaction times have appeared (e.g., Krajbich et al. 2015; Tinghög et al. 2013). They argue that reaction times in decision situations cannot be used to draw clear conclusions about the presence of intuitive or reflexive behavior, because reaction times are also influenced by other factors, such as the cognitive availability of the chosen behavior or the simplicity of the decision situation. However, this criticism only applies to a limited extent to the (Brief) Implicit Association Test, since here the within-person differences in latencies are not used as a basis for inferring intuitive or reflexive behavior, but as a measure of attitude strength.

¹⁰ The exact procedure is a bit more complicated because very long latencies (>10 seconds) are excluded from the calculation. Details can be found in Greenwald et al. (2003).

-2 and +2. In our case, a positive (negative) value for implicit racism indicates that the respondent has an implicitly more negative (more positive) attitude toward "Refugees" than toward "Germans". Similarly, a positive (negative) value for implicit populism indicates that the respondent has an implicitly more positive (negative) attitude toward "Alternatives" than toward "Establishment". The value 0 on the *D*-measure indicates an implicit indifference.

"D" für Andere

"K" für Kategorie

Deutsche
oder
gut

Thomas

Wenn Sie einen FEHLER machen, werden Sie ein X sehen. Wählen Sie dann bitte die korrekte Antwort, um fortzufahren.

Figure 1: Example of an assignment task of the BIAT.

As mentioned above, two BIATs are used in this study. Both the order of the two BIATs and the order of the individual blocks within the two BIATs were randomized. Following Sriram and Greenwald (2009), respondents also completed a practice block (attitude objects "Mammals" and "Birds") before taking the two BIATs to familiarize them with the technical procedure. The word lists of the two BIATs can be found in Table 1 in the Appendix.

Variables

The dependent variable is the intention to vote for the AfD, which was measured via a voting intention question. 9.2% of respondents stated that they would vote for the AfD.

Implicit attitudes were measured as explained above. While there is a weak tendency toward implicit racist attitudes among the respondents ($M = 0.525$, $SD = 0.655$), we find implicit populism less strongly represented ($M = -0.634$, $SD = 0.737$).

In addition to the two implicit attitudes, the explicit attitudes toward populism and racism and the disposition toward Type 1 processes are central independent variables in explaining affinity for the AfD. All three measures were collected using item batteries. Each battery consisted of several statements, whereby respondents could express their agreement with these statements on a Likert scale from 1 ("strongly disagree") to 7 ("strongly agree") (see Table 2 in the Appendix).

Explicit attitudes toward populism were operationalized using an additive index of 8 items (Cronbach's $\alpha = 0.875$, $M = 4.505$, $SD = 1.203$). Respondents had the opportunity to express their agreement with statements such as "The people often agree, but the politicians pursue completely different goals". Based on 18 statements about refugees (example: "They take away the jobs of the Germans."), the variable explicit racism results as an additive index (Cronbach's $\alpha = 0.969$, $M = 3.696$, $SD = 1.534$).

As a dispositive measure of the tendency to rely on intuitive Type 1 processes, an additive index over three statements (Cronbach's $\alpha = 0.584$, $M = 4.947$, $SD = 0.975$), which together form the short scale for measuring social desirability in terms of self-deception (cf. Winkler et al. 2006). In the data analysis, we distinguish between intuitive and reflective respondents. Intuitive respondents are individuals who score high on this measure, trust their initial intuitions, and do not tend to question them through a conscious process of reflection. Reflective respondents, on the other hand, are individuals who score low on this measure and are characterized by consciously questioning their initial intuitions.

A number of sociodemographic control variables are included in the multivariate analyses. The variables age ($M = 47.286$, $SD = 15.501$) and male gender ($M = 0.472$) are self-explanatory.¹¹ The variable migration is a dummy that takes the value 1 if the respondent's mother or father was not born in Germany ($M = 0.109$). The values for 12 respondents, who would otherwise be missing, are imputed from socio-demographic variables using a regression. The variable city is a dummy indicating whether a respondent lives in a large city or on the outskirts of a large city ($M = 0.457$). The variable West is a dummy indicating whether a respondent currently lives in West Germany ($M = 0.583$). For the respondent's highest level of education, we use three dummies to distinguish between high school diploma ("Abitur", $M = 0.529$), intermediate school ("Realschule", $M = 0.403$) and basic school ("Hauptschule", $M = 0.068$). For the employment status immediately before the Corona crisis, we use dummies to differentiate between employed ($M = 0.624$), retired ($M = 0.183$) and apprentice/student ($M = 0.131$). Due to low case numbers, we work here with a heterogeneous reference category Other ($M = 0.061$), which includes the unemployed, the disabled and domestic workers.

Net household income was openly asked. From this and from information on the composition of the household, the household equivalent income can be determined using the OECD scale. The values for 16 respondents, who would otherwise be missing, are imputed from socio-demographic variables by means of a regression. Equivalent household income is not included in the analyses directly, but through the strata derived from it. Respondents whose equivalent household income does not exceed 70% of the median equivalent household income¹² are assigned to the lower class ($M = 0.468$). Respondents whose equivalent household income is more than 70% but not more than 150% of the median equivalent

¹¹ For gender, two people chose the option to describe themselves "in a different way". Due to the small number of cases, they had to be excluded from the analyses.

¹² According to the Federal Statistical Office, in 2019 the median equivalent household income per month in Germany was (rounded to the nearest integer) 1960 euros (Federal Statistical Office 2021).

household income are assigned to the middle class ($M = 0.469$). All other respondents for whom equivalized household income is available are assigned to the upper class ($M = 0.064$).

The respondent's current or most recent occupation was collected using a drop-down menu that automatically transforms the selected occupations into the ISCO-08 classification. From this, the ISCO-88 classification can be derived using a transformation table from the International Labour Organization (ILO 1990). This allowed us to determine the SIOPS measure of occupational prestige using a procedure developed by Ganzeboom and Treiman (1996, see also Ganzeboom and Treiman 2011) and modified by Hendrickx (2004) (Stataado "ISKO") ($M = 45.705$, $SD = 9.992$). The use of the drop-down menu resulted in a high number of missing values; we imputed 520 missing values using a regression of socio-demographic variables.

Finally, we control for two general measures of cultural orientation. The religiosity variable indicates how often respondents attend religious ceremonies, such as church services ($M = 0.732$, $SD = 1.064$).¹³ General political orientation is operationalized using the left-right self-assessment ($M = 3.581$, $SD = 1.080$).¹⁴

All metric variables are z-standardized in the analyses. Out of 1102 cases, 960 cases (after imputation for migration, equivalized household income and SIOPS) have no missing values for any of the variables considered; all our analyses refer to these cases only. Almost all of the missing values occurred in the survey of implicit attitudes with the BIAT; for understandable reasons, we refrain from imputing missing values for these key independent variables.

4 Empirical findings

Bivariate relationships

As suggested by the DPP, explicit and implicit attitudes are only weakly correlated (see Figure 2). In the case of populism, the correlation coefficient is only $r = 0.089$ ($p < 0.006$); a change of four standard deviations in implicit populism is thus only accompanied by a marginal change of less than half a standard deviation in explicit populism. Implicit and explicit racism are somewhat more strongly related, but still very weak by conventional standards ($r = 0.204$, $p < 0.001$). An increase of four standard deviations in implicit racism is statistically associated with an increase of less than one standard deviation in explicit racism.

¹³ Scale: 1 "every day" to 7 "never" (negative polarity). For simplicity, we treat the variable as a metric variable.

¹⁴ Scale: 1 "left" to 7 "right". For simplicity, we treat the variable as a metric variable.

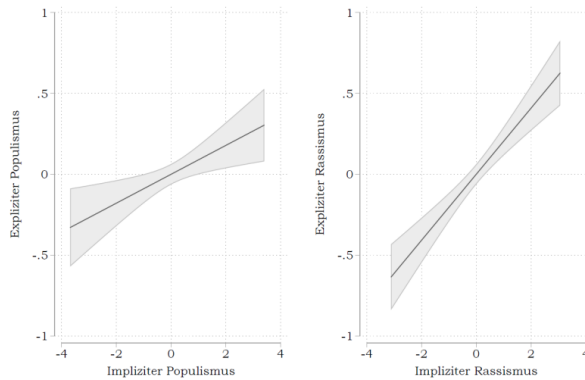


Figure 2: Correlation between implicit and explicit attitudes.

Against this background, implicit and explicit attitudes can be seen as partially independent predictors of affinity for the AfD. Figures 3 and 4 provide an overview of how the four attitudes are distributed among AfD voters and non-voters. The visual impression is confirmed by inferential statistical comparisons of means (two-tailed t-tests): AfD voters and non-voters differ significantly in all four attitudes. The smallest t-value is found for implicit racism with $t = -3.113$ and an empirical significance level of $p < 0.002$. What becomes clear is that AfD voters differ from AfD non-voters more in their explicit than in their implicit attitudes.

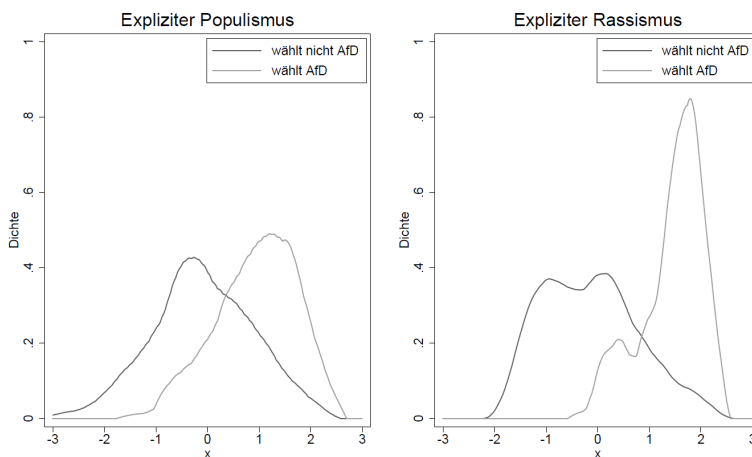


Figure 3: Explicit attitudes among AfD voters and non-voters.

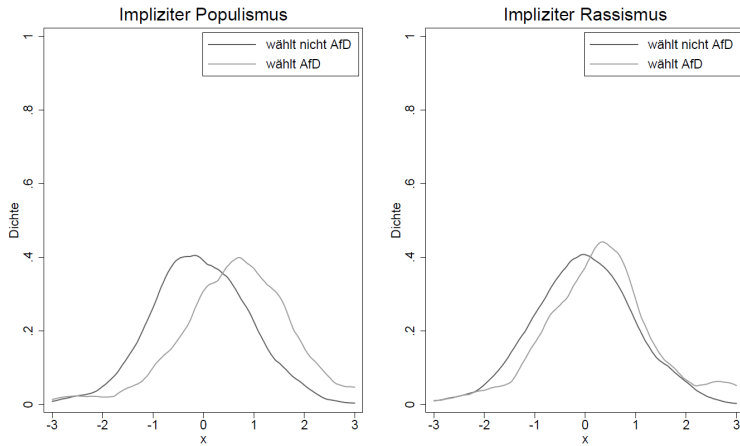


Figure 4: Implicit attitudes among AfD voters and non-voters.

Multivariate analyses

Let us now turn to testing our hypothesis in the context of multivariate models. First, we ask whether implicit attitudes are indeed more strongly reflected in explicit attitudes when the articulation of the explicit attitude takes place in the context of a Type 1 process. To this end, we estimate a regression model for each of the two explicit attitudes. The three central independent variables are the corresponding implicit attitude, our dispositive measure of tendency to Type 1 processes, and an interaction term between these two independent variables. We control for the socio-demographic variables listed in section 3 (age, gender, migration background, urban/rural, East/West, education, class, employment status, occupational prestige), for left-right self-assessment, and also for religiosity.

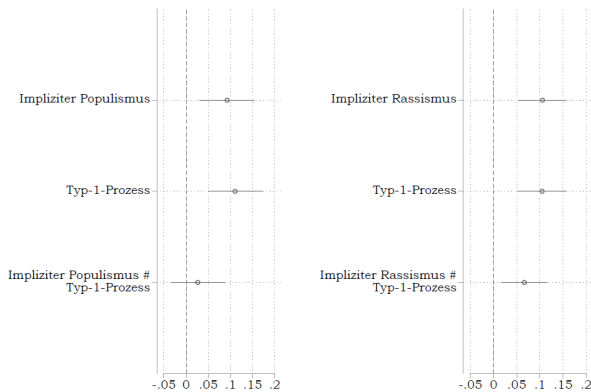


Figure 5: Explicit attitudes as a function of implicit attitudes.

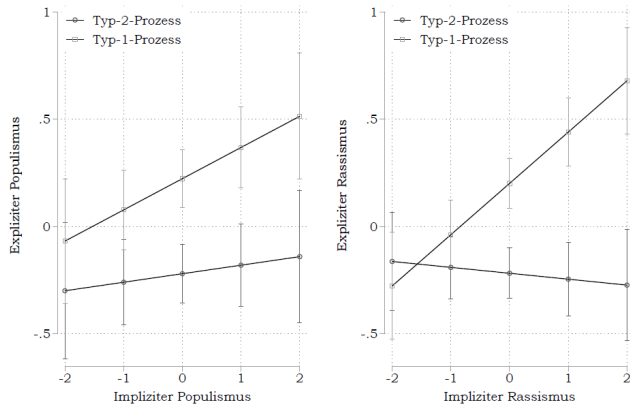


Figure 6. Interaction between implicit attitudes and intuitiveness.

Figure 5 shows that there is a descriptively positive but not significant interaction effect for populism ($p < 0.392$), while there is a positive and significant interaction effect for racism ($p < 0.006$) (see also Table 3 in the appendix). Figure 6 helps to assess the strength of the interaction effects. The straight lines labeled “Typ-1 Prozess” (Type 1 Process) reflect the predicted values for highly intuitive respondents who are two standard deviations above the mean of the dispositive measure. Similarly, the straight lines labeled “Typ-2 Prozess” (Type 2 Process) report the predicted values for highly reflective respondents, who are two standard deviations below the mean of this measure. Looking at populism, we find an interaction effect that is not particularly large in terms of effect size; the straight line labeled “Typ-1 Prozess” has about twice the slope of the straight line labeled “Typ-2 Prozess”. But double of almost nothing is still little. In the case of racism, however, there is a rather strong interaction effect: In fact, among more reflective respondents, implicit racism has no effect on explicit racism, while among more intuitive respondents there is a quite noticeable relationship. Overall, for both attitudes, i.e., populism and racism, we interpret these findings as clearly positive evidence for the validity of the principle of catalyzation.

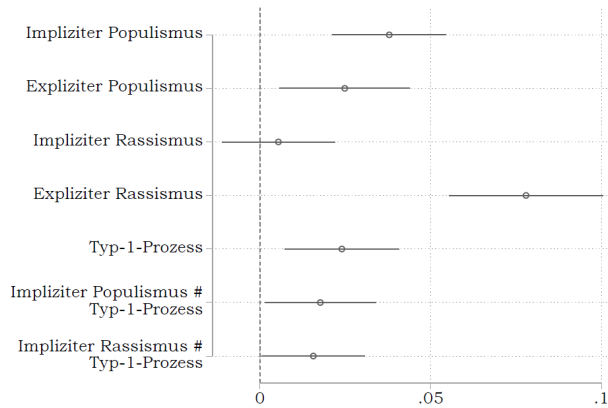


Figure 7: Affinity to the AfD as a function of implicit and explicit attitudes.

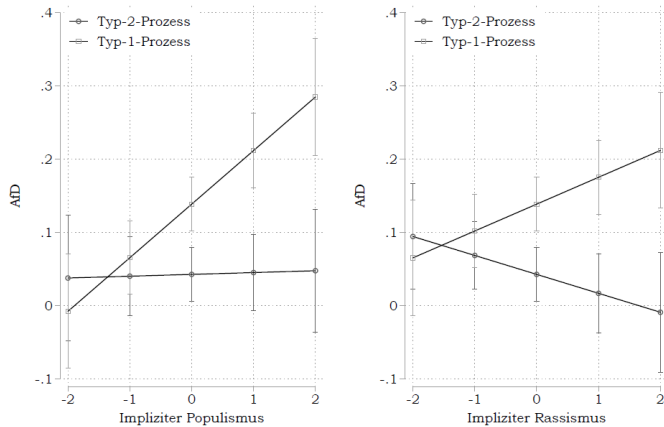


Figure 8: Interaction between implicit attitudes and intuitiveness.

Let us now turn to the consideration of voting intentions for the AfD. We estimate a model that includes all four attitudes into account simultaneously as independent variables, as well as the measure of intuitiveness and the two interaction effects between this measure and the two implicit attitudes. In addition, we again use the control variables mentioned above. Given that our theoretical interest is focused on interaction effects and that these are difficult to test for in the context of logit or probit models (Ai and Norton 2003), we follow the recommended practice in econometrics and increasingly also in sociology and resort to the linear probability model (Breen et al. 2018).

We find positive and significant interaction effects for both implicit populism and implicit racism (see Figure 7 and Table 4 in the appendix).¹⁵ Figure 8 shows that implicit populism has no effect on affinity for the AfD among highly reflective respondents, while there is a positive correlation among highly intuitive respondents. Implicit racism also has a significant positive influence on affinity for the AfD among highly intuitive respondents (see Figure 8). However, the small and insignificant main effect of implicit racism also means that there is even a negative correlation between implicit racism and affinity for the AfD among highly reflective respondents. A possible explanation for this - at first sight irritating - finding could be repression in the sense of Freud (1911), which can be reconstructed in the DPP framework as a conflict between implicit and explicit orientations (cf. Wilson et al. 2000). This means that highly reflective respondents with strong implicit racist attitudes might not vote for the AfD in order to avoid becoming aware of their implicit attitudes. This is, of course, only a speculative post-hoc interpretation of a finding that should not be exaggerated. Indeed, an additional analysis, looking only at respondents with an above-average tendency to reflect, shows that this negative correlation between implicit racism and affinity for the AfD is not significant ($b = -0.011, p < 0.300$).

Incidentally, the effect sizes of the implicit attitudes are remarkable. The model on which Figures 7 and 8 are based shows a coefficient for left-right self-assessment of 0.048 (see Table 4 in the appendix). Thus, *ceteris paribus*, a one standard deviation increase in one's self-assessment as politically right-wing leads to a 4.8 percentage point increase in the probability of voting for the AfD. According to the model, East Germans have a 4.6% increased probability of voting for the AfD. Both the left-right self-assessment and the East/West dummy are known to be relatively strong predictors in relevant research on the AfD (cf. Hambauer and Mays 2018; Lengfeld 2017). However, among respondents whose intuitiveness is one standard deviation above the mean, implicit populism has a stronger effect: a one standard deviation increase in implicit populism leads to a 5.8% increase in the probability of voting for the AfD. In fact, implicit populism has a stronger influence than explicit populism and can be interpreted as the second strongest predictor overall after explicit racism. We interpret these findings on the influence of implicit attitudes on the affinity to the AfD as a clear confirmation of the catalyzation principle.

5 Discussion

In this paper, the distinction between implicit and explicit attitudes was taken up on the basis of the DPP and it was argued that the DPP formulates a general condition for implicit attitudes to be reflected more strongly in explicit attitudes on the one hand, and in overt behavior on

¹⁵ Figure 7 is based on a model without robust standard errors. Since the assumption of homoscedasticity is violated in the linear probability model, one also finds the recommendation to work with robust standard errors. Here, weakly significant interaction effects emerge (Implicit Populism # Type 1 Process: $p < 0.065$, Implicit Racism # Type-1 Process: $p < 0.069$).

the other. According to the principle of catalyzation, the influence of implicit attitudes increases when the explication or action occurs in the context of Type 1 processes. These action-theoretical ideas were applied in the context of explaining affinity for the AfD. In this exploratory study, it was shown that, on the one hand, implicit attitudes toward populism and racism are only weakly related to corresponding explicit attitudes. Furthermore, both types of attitudes are significant and strong predictors of voting intentions for the AfD. Finally, the catalysis principle is also valid, as more reflective respondents show a weaker correlation between implicit attitudes and explicit attitudes and also between implicit attitudes and affinity for the AfD than respondents who rely more on their intuition.

These findings and the underlying action-theoretical argumentation can provide fruitful impulses for electoral research and, in particular, for empirical research on right-wing populist or extremist tendencies. In general, the political science and sociological literature considers both objectively given and subjectively perceived socioeconomic status (Brug et al. 2000; Lubbers et al. 2002; Arzheimer and Carter 2006) and cultural orientations such as racism, xenophobia, authoritarianism, and Euroscepticism (Kitschelt 1995; Ivarsflaten 2005; Werts et al. 2013) as predictors of right-wing populist tendencies. However, cultural orientations are almost exclusively measured by explicit attitudinal measures. Given our findings that implicit and explicit attitudes are only closely related under certain conditions, and that implicit attitudes can influence voting intentions even when explicit attitudes are controlled for, it can be concluded that central determinants of individual voting behavior have been overlooked in the most election research. The thesis that implicit attitudes cannot be neglected when explaining right-wing populist tendencies is also relevant for the sociological debate on the AfD. Here, following contributions by Ronald Inglehart and Pippa Norris (Inglehart and Norris 2017, 2018), the question of whether socioeconomic factors (Economic Insecurity Hypothesis) or cultural orientations (Cultural Backlash Hypothesis) determine affinity for the AfD is controversial (Lengfeld 2017; Lux 2018; Tutić and von Hermanni 2018; Rippl and Seipel 2018; Lengfeld and Dilger 2018). A well-founded discussion of this question requires mediation analyses, because in addition to the direct effects of socioeconomic position, which also affect affinity for the AfD when controlling for cultural orientations, the indirect effects mediated by differences in these cultural orientations must also be taken into account (cf. Lengfeld and Dilger 2018). Again, from an action-theoretical perspective, the contributions presented so far suffer from the fact that cultural orientations are considered exclusively by means of explicit measures. The action-theoretical perspective outlined in this article leads to additional challenges for empirical social research: multiple mediation analyses are necessary because status positions influence voting intentions not only directly, but also indirectly via implicit and explicit attitudes, which in turn are moderated by the disposition to act intuitively or reflectively. It can be assumed that socioeconomic status affects not only the implicit and explicit attitudes of actors, but also their disposition toward more intuitive or reflective behavior (cf. Brett and Miles 2021). In light of our findings on the relevance of implicit attitudes for AfD voting intentions, multiple mediation analyses that take into account the aforementioned considerations are a necessary and logical endeavor to advance the debate on the relative empirical validity of the Economic Insecurity Hypothesis and the Cultural Backlash Hypothesis. Unfortunately, our data lend themselves to such analyses only to a limited extent, primarily because the implicit measures of populism and racism cover only a very limited spectrum of implicit cultural orientations.

At this point, we would like to address and discuss another finding on the main effects of our measure of intuitiveness, which is not the focus of this paper but should be of interest nonetheless. We find a consistent pattern that the propensity for intuitiveness is associated with agreeing with explicitly populist and racist statements and also with being more likely to vote for the AfD (see Figures 5 and 7). At first glance, this finding contradicts the common hypothesis that intuitiveness and spontaneity favor prosocial behavior (Rand et al. 2012; Rand 2016; Rand et al. 2014). In this regard, we would like to elaborate on three aspects. First, our findings seem plausible because our measure of intuitiveness focuses on the tendency to rely on first impressions and “common sense”, and this tendency is served by populist parties (such as the determined rejection of “elitist” experts). Moreover, the thesis of intuitive prosociality is also not uncontroversial in interdisciplinary research (Bouwmeester et al. 2017; Verkoeijen and Bouwmeester 2014). According to the DPP, implicit attitudes should have a stronger influence on behavior when that behavior occurs as part of a Type 1 process. Based on this catalyzation principle, one only arrives at the thesis of intuitive prosociality if one assumes a general implicit orientation of humans; and it is precisely this assumption that is empirically questionable. Finally, our results fit well into the interdisciplinary discourse on DPP. Empirical evidence for intuitive prosociality comes mainly from studies of economic games whose payoff structure is such that individual payoffs and social welfare come into conflict (logic: “I” versus “group”). However, the political narrative of right-wing populist parties such as the AfD focuses not on the conflict between the individual and society, but on the conflict between social aggregates or groups (logic: “us” versus “the others”): On the one hand, on the conflict between “the people” and “the elite” and, on the other hand, on the conflict between “Germans” and “foreigners”. With Greene (2013) it can now be argued that the human species, as an evolutionary adaptation, has developed the tendency to intuitively act prosocially in situations of conflict between self-interest and group interest, but that this prosociality is always limited to one's own group. In situations structured according to the logic of “us” versus “the others”, intuitive prosociality toward one's own group leads to intuitive antisociality toward the others.

Let us now turn to the limitations of this paper. In section 2, we pointed out that from a theoretical perspective, the relationship between explicit attitudes and behavior is difficult to grasp. Depending on the mode in which the articulation of the explicit attitude and the behavior comes about, four cases can be distinguished. The reliance on a dispositive measure of Type 1/Type 2 processes means that we can only compare cases in which both the explication of attitudes and the voting intention arose as part of a Type 1 process with cases in which neither the explication nor the behavior was processed automatically-spontaneously. For this reason, the present study is not very instructive regarding the influence of explicit attitudes. Future research should use techniques such as time pressure or cognitive overload to focus empirically on the other two cases in which either the explicit attitude or the behavior is automatic-spontaneous. In addition, it would be interesting to investigate whether our measure of Type 1/Type 2 processes proves valid in other contexts and how it relates to alternative measures, such as attitude accessibility (Mayerl 2010) or attitude anchoring (Stocké 2004).

With regard to the empirical example, the main limitation of the present study is certainly that we cannot claim to be representative of the German population and that we are working with a cross sectional study. Interesting hypotheses about the relative importance of

socioeconomic factors and cultural orientations or about the prevalence of implicit and explicit attitudes in the German population over time cannot be tested on this data basis. From our point of view, it would be a very worthwhile undertaking to set up a representative panel that includes implicit cultural orientations in addition to the standard variables associated with research on right-wing populist tendencies.

Literature

- Ai, Chunrong and Edward Norton. 2003. Interaction Terms in Logit and Probit Models. *Economics Letters* 80: 123–129.
- Arzheimer, Kai and Elisabeth Carter. 2006. Political Opportunity Structures and Right-Wing Extremist Party Success. *European Journal of Political Research* 45: 419–443.
- Beier, Harald. 2016. Wie wirken 'Subkulturen der Gewalt'? Das Zusammenspiel von Internalisierung und Verbreitung gewaltlegitimierender Normen in der Erklärung von Jugendgewalt. *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 68: 457–485.
- Bourdieu, Pierre. 1990. *The Logic of Practice*. Stanford: Stanford University Press.
- Bouwmeester, Samantha, Peter P. J. L. Verkoeijen, Balazs Aczel et al. 2017. Registered Replication Report: Rand, Greene, and Nowak (2012). *Perspectives on Psychological Science* 12: 527–542.
- Breen, Richard, Kristian B. Karlson and Anders Holm. 2018. Interpreting and Understanding Logits, Probits, and Other Nonlinear Probability Models. *Annual Review of Sociology* 44: 39–54.
- Brett, Gordon and Andrew Miles. 2021. Who Thinks How? Social Patterns in Reliance on Automatic and Deliberate Cognition. *Sociological Science* 8: 96–118.
- Brug, Wouter van der, Meindert Fennema and Jean Tillie. 2000. Anti-Immigrant Parties in Europe: Ideological or Protest Vote? *European Journal of Political Research* 37: 77–102.
- Decker, Oliver, Johannes Kiess, Julia Schuler, Barbara Handke, Gert Pickel and Elmar Brähler. 2020. 2. Die Leipziger Autoritarismus Studie 2020: Methode, Ergebnisse und Langzeitverlauf. In *Autoritäre Dynamiken*, 27–88. Gießen: Psychosozial-Verlag.
- DiMaggio, Paul. 1997. Culture and Cognition. *Annual Review of Sociology* 23: 263–287.
- Durkheim, Emile. 1997. *The Division of Labour in Society*. New York: Free Press.
- Eagly, Alice H. and Shelly Chaiken. 1993. *The Psychology of Attitudes*. San Diego: Harcourt Brace Janovich.
- Esser, Hartmut. 1996. Die Definition der Situation. *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 48: 1–34.
- Esser, Hartmut and Clemens Kroneberg. 2020. Das Modell der Frame-Selektion. In *Rational Choice*, Hrsg. Andreas Tutić, 308–324. Berlin: DeGruyter.

- Evans, Jonathan. 2010. *Thinking Twice: Two Minds in One Brain*. Oxford: Oxford University Press.
- Evans, Jonathan and Keith E. Stanovich. 2013. Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives on Psychological Science* 8: 223–241.
- Fazio, Russel H. 1995. Attitudes as Object-Evaluation Associations: Determinants, Consequences, and Correlates of Attitude Accessibility. In *Attitude Strength. Antecedents and Consequences*, Hrsg. R. E. Petty and J. A. Krosnick, 247–282. New Jersey: Lawrence Erlbaum Associates.
- Fazio, Russell H., Joni R. Jackson, Bridget C. Dunton and Carol J. Williams. 1995. Variability in Automatic Activation as an Unobtrusive Measure of Racial Attitudes: A Bona Fide Pipeline? *Journal of Personality and Social Psychology* 69: 1013–1027.
- Fazio, Russell H. and Michael A. Olson. 2003. Implicit Measures in Social Cognition: Their Meaning and Use. *Annual Review of Psychology* 54: 297–327.
- Freud, Sigmund. 1911. Psychoanalytische Bemerkungen über einen autobiographisch beschriebenen Fall von Paranoia (Dementia paranoides). *Jahrbuch Für Psychoanalytische und Psychopathologische Forschungen* 3: 9–68.
- Ganzeboom, Harry B. and Donald J. Treiman. 1996. Internationally Comparable Measures of Occupational Status for the 1988 International Standard Classification of Occupations. *Social Science Research* 25: 201–239.
- Ganzeboom, Harry B. and Donald J. Treiman. 2011. International Stratification and Mobility File: Conversion Tools. <http://www.harryganzeboom.nl/ismf/index.html>.
- Giddens, Anthony. 1984. *The Constitution of Society: Outline of the Theory of Structuration*. Cambridge: Polity Press.
- Greene, Joshua. 2013. *Moral Tribes. Emotion, Reason, and the Gap between Us and Them*. New York: Penguin Books.
- Greene, Joshua D., R. Brian Sommerville, Leigh E. Nystrom, John M. Darley and Jonathan D. Cohen. 2001. An fMRI Investigation of Emotional Engagement in Moral Judgment. *Science* 293: 2105–8.
- Greenwald, Anthony G., Debbie E. McGhee and Jordan L. Schwartz. 1998. Measuring Individual Differences in Implicit Cognition: The Implicit Association Test. *Journal of Personality and Social Psychology* 74: 1464–1480.
- Greenwald, Anthony G., Brian A. Nosek and Mahzarin R. Banaji. 2003. Understanding and Using the Implicit Association Test: I. An Improved Scoring Algorithm. *Journal of Personality and Social Psychology* 85: 197–216.
- Hambauer, Verena and Anja Mays. 2018. Wer wählt die AfD?–Ein Vergleich der Sozialstruktur, politischen Einstellungen und Einstellungen zu Flüchtlingen zwischen AfD-

WählerInnen und der WählerInnen der anderen Parteien. *Zeitschrift für Vergleichende Politikwissenschaft* 12: 133–154.

Hendrickx, John. 2004. ISKO: Stata Module to Recode 4 Digit ISCO-88 Occupational Codes.

Homans, George C. 1974. *Social Behavior: Its Elementary Forms*. San Diego: Harcourt Brace Jovanovich.

ILO. 1990. International Standard Classification of Occupations (ISCO-88). Geneva: International Labour Office.

Inglehart, Ronald and Pippa Norris. 2017. Trump and the Populist Authoritarian Parties: The Silent Revolution in Reverse. *Perspectives on Politics* 15: 443–454.

Ivarsflaten, Elisabeth. 2005. The Vulnerable Populist Right Parties: No Economic Realignment Fuelling their Electoral Success. *European Journal of Political Research* 44: 465–492.

Kahneman, Daniel. 2011. *Thinking, Fast and Slow*. London: Penguin Books.

Kitschelt, Herbert. 1995. *The Radical Right in Western Europe. A Comparative Analysis*. Ithaca: Cornell University Press.

Krajbich, Ian, Bjorn Bartling, Todd Hare and Ernst Fehr. 2015. Rethinking Fast and Slow Based on a Critique of Reaction-Time Reverse Inference. *Nature Communications* 6: 7455.

Kroneberg, Clemens. 2005. Die Definition der Situation und die variable Rationalität der Akteure. Ein allgemeines Modell des Handelns. *Zeitschrift für Soziologie* 34: 344–363.

Kroneberg, Clemens. 2007. Wertrationalität und das Modell der Frame-Selektion. *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 59: 215–239.

Kroneberg, Clemens. 2011. *Die Erklärung sozialen Handelns*. Wiesbaden: VS Verlag.

Kroneberg, Clemens, Meir Yaish and Volker Stocké. 2010. Norms and Rationality in Electoral Participation and in the Rescue of Jews in WWII: An Application of the Model of Frame Selection. *Rationality and Society* 22: 3–36.

Lengfeld, Holger. 2017. Die 'Alternative für Deutschland': Eine Partei für Modernisierungsverlierer? *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 69: 209–232.

Lengfeld, Holger and Clara Dilger. 2018. Kulturelle und ökonomische Bedrohung. Eine Analyse der Ursachen der Parteidentifikation mit der Alternative für Deutschland mit dem Sozio-ökonomischen Panel 2016. *Zeitschrift für Soziologie* 47: 181–199.

Lindenberg, Siegwart. 2013. Social Rationality, Self-Regulation, and Well-Being: The Regulatory Significance of Needs, Goals, and the Self. In *The Handbook of Rational Choice Social Research*, Hrsg. R. Wittek, T. Snijders and V. Nee, 72–112. Palo Alto: Stanford University Press.

Lubbers, Marcel, Merove Gijsberts and Peer Scheepers. 2002. Extreme Right-Wing Voting in Western Europe. *European Journal of Political Research* 41: 345–378.

Lux, Thomas. 2018. Die AfD und die unteren Statuslagen. Eine Forschungsnotiz zu Holger Lengfelds Studie 'Alternative für Deutschland': Eine Partei für Modernisierungsverlierer? *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 70: 255–273.

Martin, John Levi. 2010. Life's a Beach but You're an Ant, and other Unwelcome News for the Sociology of Culture. *Poetics* 38: 229–244.

Mayerl, Jochen. 2010. Die Low-Cost-Hypothese ist nicht genug. *Zeitschrift für Soziologie* 39: 38–59.

McClure, Samuel M., David I. Laibson, George Loewenstein and Jonathan D. Cohen. 2004. Separate Neural Systems Value Immediate and Delayed Monetary Rewards. *Science* 306: 503–507.

Mudde, Cas. 2007. *Populist Radical Right Parties in Europe*. Cambridge: Cambridge University Press.

Norris, Pippa and Ronald Inglehart. 2018. *Cultural Backlash. Trump, Brexit, and the Rise of Authoritarian Populism*. New York: Cambridge University Press.

Nosek, Brian A., Yoav Bar-Anan, Natarajan Sriram, Jordan Axt and Anthony G. Greenwald. 2014. Understanding and Using the Brief Implicit Association Test: Recommended Scoring Procedures. *PLoS ONE* 9: e110938.

Parsons, Talcott. 1937. *The Structure of Social Action*. Glencoe: Free Press.

Rand, David G. 2016. Cooperation, Fast and Slow: Meta-Analytic Evidence for a Theory of Social Heuristics and Self-Interested Deliberation. *Psychological Science* 27: 1192–1206.

Rand, David G., Joshua D. Greene and Martin A. Nowak. 2012. Spontaneous Giving and Calculated Greed. *Nature* 489: 427–430.

Rand, David G. and Gordon T. Kraft-Todd. 2014. Reflection Does Not Undermine Self-Interested Prosociality. *Frontiers in Behavioral Neuroscience* 8: 300.

Rand, David G., Alexander Peysakhovich, Gordon T. Kraft-Todd, George E. Newman, Owen Wurzbacher, Martin A. Nowak and Joshua D. Greene. 2014. Social Heuristics Shape Intuitive Cooperation. *Nature Communications* 5: 3677.

Reckwitz, Andreas. 2003. Grundelemente einer Theorie sozialer Praktiken. Eine sozialtheoretische Perspektive. *Zeitschrift für Soziologie* 32: 282–301.

Rippl, Susanne and Christian Seipel. 2018. Modernisierungsverlierer, Cultural Backlash, Postdemokratie - Was erklärt rechtspopulistische Orientierungen? *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 70: 237–254.

Rubinstein, Ariel. 1998. *Modeling Bounded Rationality*. Cambridge: MIT Press.

Rubinstein, Ariel. 2013. Response Time and Decision Making. A 'Free' Experimental Study. *Judgement and Decision Making* 8: 540–551.

Rydell, Robert J. and Allen R. McConnell. 2006. Understanding Implicit and Explicit Attitude Change: A Systems of Reasoning Analysis. *Journal of Personality and Social Psychology* 91: 995–1008.

Sriram, Natarajan and Anthony G. Greenwald. 2009. The Brief Implicit Association Test. *Experimental Psychology* 56: 283–294.

Sriram, Natarajan, Brian A. Nosek and Anthony Greenwald. 2006. Scale Invariant Contrasts of Response Latency Distributions. *Available at SSRN 2213910*.

Stanovich, Keith E. 2011. *Rationality and the Reflective Mind*. Oxford: Oxford University Press.

Statista. 2020. Rechtsextremismus: Rechtsextremistische Straftaten und Gewalttaten bis 2019. de.statista.com/statistik/daten/studie/4032/umfrage/rechtsextremismus-undfremdenfeindlichkeit-in-deutschland/ (Zugegriffen am 2. April 2021).

Statistisches Bundesamt. 2021. Lebensbedingungen und Armutsgefährdung. Einkommensverteilung (Nettoäquivalenzeinkommen) in Deutschland. destatis.de/DE/Themen/Gesellschaft-Umwelt/Einkommen-Konsum-Lebensbedingungen/Lebensbedingungen-Armutsgefahrdung/Tabellen/einkommensverteilung-silc.html (Zugegriffen am 3. Mai 2021).

Stocké, Volker. 2004. Entstehungsbedingungen von Antwortverzerrungen durch soziale Erwünschtheit. *Zeitschrift für Soziologie* 33: 303–320.

Strack, Fritz and Roland Deutsch. 2004. Reflective and Impulsive Determinants of Social Behavior. *Personality and Social Psychology Review* 8: 220–247.

Strotz, Robert H. 1955. Myopia and Inconsistency in Dynamic Utility Maximization. *Review of Economic Studies* 23: 165–180.

Swidler, Ann. 1986. Culture in Action: Symbols and Strategies. *American Sociological Review* 51: 273–286.

Tinghög, Gustav, David Andersson, Caroline Bonn, Harald Böttiger, Camilla Josephson, Gustaf Lundgren, Daniel Västfjäll, Michael Kirchler and Magnus Johannesson. 2013. Intuition and Cooperation Reconsidered. *Nature* 498: E1–E2.

Tutić, Andreas. 2015a. Revealed Norm Obedience. *Social Choice and Welfare* 44: 301–318.

Tutić, Andreas. 2015b. Warum denn eigentlich nicht? Zur Axiomatisierung soziologischer Handlungstheorie. *Zeitschrift für Soziologie* 44: 83–95.

Tutić, Andreas and Hagen von Hermanni. 2018. Sozioökonomischer Status, Deprivation und die Affinität zur AfD - Eine Forschungsnotiz. *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 70: 275–294.

Urban, Dieter and Jochen Mayerl. 2007. Antwortlatenzzeiten in der survey-basierten Verhaltensforschung. *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 59: 692–713.

Vaisey, Stephen. 2009. Motivation and Justification: A Dual-Process Model of Culture in Action. *American Journal of Sociology* 114: 1675–1715.

Vaisey, Stephen. 2014. Is Interviewing Compatible with the Dual-Process Model of Culture. *American Journal of Cultural Sociology* 2: 150–158.

Verkoeijen, Peter P. J. L. and Samantha Bouwmeester. 2014. Does Intuition Cause Cooperation? *PLOS ONE* 9: e96654.

Weber, Max. 2002. *Wirtschaft Und Gesellschaft*. Tübingen: Mohr Siebeck.

Werts, Han, Peer Scheepers and Marcel Lubbers. 2013. Euro-Scepticism and Radical Right-Wing Voting in Europe, 2002-2008: Social Cleavages, Socio-Political Attitudes and Contextual Characteristics Determining Voting for the Radical Right. *European Union Politics* 14: 183–205.

Wilson, Timothy D. 2002. *Strangers to Ourselves. Discovering the Adaptive Unconscious*. Cambridge: Belknap Press.

Wilson, Timothy D., Samuel Lindsey and Tonya Y. Schooler. 2000. A Model of Dual Attitudes. *Psychological Review* 107: 101–126.

Winkler, Niels, Martin Kroh and Martin Spiess. 2006. Entwicklung einer deutschen Kurzskaala zur zweidimensionalen Messung von sozialer Erwünschtheit. *DIW Berlin, German Institute for Economic Research, Discussion Papers of DIW Berlin*.

Zieck, Andreas, Beate Küpper and Wilhelm Berghan. 2019. *Verlorene Mitte. Feindselige Zustände. Rechtsextreme Einstellungen in Deutschland 2018/19*. Bonn: Dietz.

Appendix

Table 1: Word lists for the Brief Implicit Association Test. With regard to the first names of Germans and refugees, we were guided by lists of the most common first names in Germany and in Arabic-speaking countries, since at the time of the survey (2020) the majority of refugees came from countries such as Syria or Iraq.

Category	Associated words
Attribute	
Good	nice, heat, love, friend
Badly	mean, cold, hate, enemy
Attribute object	
German	Paul, Thomas, Marie, Sophia
Refugees	Enis, Ali, Fatima, Aischa
Established	Angela Merkel, CDU, Public broadcasting, Bill Gates
Alternative	Björn Höcke, AfD, Trump, Alternative media

Table 2. Items for explicit racism, explicit populism, and disposition toward Type 1 processes.

Explicit racism

Their presence in Germany leads to problems on the housing market.

They take jobs away from the Germans.

They are a burden on the social safety net.

They will make a positive contribution to Germany's economic development. +

The current number of refugees is a threat to prosperity in Germany.

They are treated better than Germans in many areas of life.

They cause social cohesion to be lost.

Their presence often causes problems in the neighborhood where they live.

The many refugee children in school prevent the German children from receiving a good education.

They had better adapt their lifestyle to that of the Germans.

With so many refugees in Germany, people increasingly feel like strangers in their own country.

They are an enrichment for the culture in Germany. +

They make Germany more tolerant and open to the world. +

They will enrich Germany culturally in the long term. +

They increase crime in Germany.

They increase the risk of terrorist attacks.

If housing becomes scarce, refugees living in Germany should be sent back to their home countries.

There are too many refugees living in Germany.

Explicit populism

The people often agree, but the politicians pursue quite different goals.

I would rather be represented by an ordinary citizen than a career politician.

Political parties are only interested in voters' votes, not their opinions.

The differences in political views between the elite and the people are greater than the differences within the people.

Important political decisions should not be decided by politicians but by the people through referendums.

Politicians in Germany must follow the will of the people.

In principle, the people of Germany agree on what should happen in politics.

What is called "compromise" in politics is really just selling one's principles.

Disposition to Type 1 processes

I always know exactly why I like something.

My first impression of people usually turns out to be correct.

I am often unsure of my judgment. +

+ Reverse coded items.

Table 3: Linear regression models. Dependent variable: explicit attitudes. (1) Explicit populism, (2) Explicit racism. Items for explicit racism, explicit populism, and disposition toward Type 1 processes.

<i>Explicit attitudes</i>	(1)		(2)			
	Racism		Populism			
	b	SE	b	SE		
Implicit populism	0.093	**	0.031			
Implicit racism				0.106	***	0.026
Type 1 process	0.111	***	0.031	0.105	***	0.027
Imp. populism × Type 1 process	0.026		0.031			
Imp. racism × Type 1 process				0.067	**	0.024
Age	0.084	+	0.048	0.058		0.041
Male	-0.121	+	0.062	-0.083		0.053
Migration	0.180	+	0.098	-0.013		0.083
City	-0.035		0.062	0.039		0.053
Education (ref.: Secondary school)						
High school	-0.216	**	0.072	-0.134	*	0.062
Basic school	0.183		0.128	0.144		0.109
Class (reference: middle class)						
Upper class	-0.087		0.127	-0.126		0.109
Lower class	0.122	+	0.067	0.106	+	0.057
Employment status (ref.: other)						
Employed	-0.112		0.130	0.094		0.111
Trainees/Students	-0.233		0.167	-0.178		0.143
Pensioners	-0.054		0.152	0.124		0.130
Professional prestige	-0.072	*	0.034	-0.097	***	0.029
West Germany	-0.286	***	0.063	-0.238	***	0.054
Religiosity	-0.021		0.031	-0.018		0.026
Left-Right self-assessment	0.148	***	0.031	0.483	***	0.027
Constant	0.382	**	0.142	0.115		0.122
Observations	960		960			
R^2	0.160		0.388			

+ $p < 0.1$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Table 4: Linear probability models. Dependent variable: AfD voting intention. (3) Normal standard errors, (4) Robust standard errors.

<i>Voting intentions for the AfD</i>	(3)		(4)	
	<i>b</i>	SE	<i>b</i>	SE
Implicit populism	0.038	*** 0.008	0.038	*** 0.009
Explicit populism	0.025	* 0.010	0.025	** 0.009
Implicit racism	0.005	0.008	0.005	0.009
Explicit racism	0.078	*** 0.011	0.078	*** 0.012
Type 1 process	0.024	** 0.008	0.024	** 0.009
Imp. populism × Type 1 process	0.018	* 0.008	0.018	+ 0.010
Imp. racism × Type 1 process	0.016	* 0.008	0.016	+ 0.009
Age	−0.006	0.013	−0.006	0.014
Male	0.033	* 0.017	0.033	* 0.017
Migration	−0.023	0.026	−0.023	0.024
City	0.023	0.017	0.023	0.017
Education (ref.: secondary school)				
High school	−0.012	0.019	−0.012	0.019
Basic school	−0.014	0.034	−0.011	0.039
Class (reference: middle class)				
Upper class	0.028	0.034	0.028	0.035
Lower class	0.014	0.018	0.014	0.019
Employment status (ref.: other)				
Employed	0.014	0.035	0.014	0.035
Trainees/Students	0.004	0.045	0.004	0.042
Pensioners	0.014	0.041	0.014	0.041
Professional prestige	0.000	0.009	0.000	0.007
West Germany	−0.046	** 0.017	−0.046	** 0.017
Religiosity	−0.005	0.008	−0.005	0.009
Left-Right self-assessment	0.048	*** 0.010	0.048	*** 0.010
Constant	0.080	* 0.039	0.080	* 0.038
Observations	960		960	
R^2	0.278		0.278	

+ $p < 0.1$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$



Graphic design: Communication Division, UIB / Print: Skjipes Kommunikasjon AS



uib.no

ISBN: 9788230858974 (print)
9788230840597 (PDF)