Dissertation summary of

Towards a comprehensive understanding of rapid instruction-based learning: neural and behavioral insights from systematically investigating rule identity conflicts

by M.Sc. Alexander Willy Baumann

Instruction-based learning (IBL) – the ability to implement novel stimulus-response (S-R) rules immediately and accurately after a single instruction – is a fundamental human cognitive capacity. Despite considerable progress in understanding its neural and behavioral underpinnings, key questions remain regarding the mechanisms that enable rule transformation from an initial action-detached into an actionable format. This is partly due to conceptual imprecision as well as to the inherent problem of objectively distinguishing between these two functional states at the level of a single rule. Therefore, the central idea expressed in this thesis is to tie functional states to S-R identity. The first research question explores whether and how this could be done. Building on this, the second research question addresses the mechanistic functioning of the rule transformation process. Finally, the third research question is dedicated to whether instructional content continues to impact behavior beyond first-time rule execution.

The present thesis addresses these questions across four empirical studies, of which two were purely behavioral and two utilized functional magnetic resonance imaging. It does so by varying instructional properties within the same task architecture and focusing especially on emerging conflicts at the level of individual S-R identities.

Study I examined functional connectivity changes induced by instructional load during rule implementation. Behaviorally, implementation performance was continuously impeded by high instructional load. This was due to an ongoing conflict between the instructional content and a self-generated rule that emerged as a result of initial implementation failure. Imaging results revealed opposing coupling profiles of lateral prefrontal cortex (PFC) regions depending on whether instructional load remained within or exceeded working memory capacity limits. The functional coupling profile of the left ventrolateral PFC was highly specific and implied relatively stable representation of instructional content that is selectively accessed by different cognitive systems – with specific couplings predicting individual learning success.

Study II compared novel (i.e., initial) learning with instructed reversal learning. The results provided behavioral evidence for proactive interference through initially learned rules on implementation performance of following instructed rule reversal. This interference was due to a durable, long(er)-term memory trace formed after as few as four implementation trials and a

memory trace formed on the basis of mere instructions. Interestingly, interference effects were not limited to first-time implementation but persisted across the entire implementation phase.

Study III introduced negative instructions – specifying the not to be executed response rather than the to be executed one – as a principled means to create objectively distinct instruction-related and implementation-related S-R identities related to a single rule. Relative to ad-hoc negation of previously specified, neutral S-R associations, it was shown that rule implementation was more efficient following negative instructions. Compared to standard, 'positive' instructions specifying affirmative rules, however, negative instructions were associated with persistent implementation costs. These costs could be attributed to the continuous presence of the instruction-related S-R association – the integrity of which was quantified by means of a recognition test. Importantly, and as evidenced by a decreasing tendency to switch between equally correct response options, this was accompanied by a gradual consolidation of one distinct implementation-related S-R association.

In study IV, a multivariate pattern approach sensitive to individual rule identities was employed in order to characterize representational dynamics following negative instructions. At this neural level there was no evidence for a gradual increase in identity-specific rule coding strength across repeated rule implementation as could have been expected based on the results of study III. Instead, stable identity-specific representational patterns in the left (ventro-)lateral PFC were present already at instruction encoding and were selectively linked to instruction-related S-R integrity in the negative condition. At the same time, evidence for covert response preparation in the motor cortex – present for positive instructions – was absent following negative instructions, indicating distinct proactive transformation processes between instruction types.

With respect to the first research question of interest, the individual study results suggest that tying functional states to distinct S-R identities is possible, for example by utilizing negative instructions – an approach that should be further pursued and refined. Regarding the second research question, study results imply that rule transformation is not achieved by representational change within one brain region but rather through emergence of relatively stable representations that are flexibly accessed – here the left ventrolateral PFC seems to play a key role. Relatedly, and of relevance for the third research question, the cognitive system seems to draw on representations or memory traces of instructional content even well beyond first-time implementation. Together, these findings highlight the informative value of rule identity conflicts and the relevance of S-R identity tagging as a means to objectivize functional states in rapid learning. This opens up a new perspective in research on instruction-induced rule transformation and needs to be considered in future works in order to arrive at a truly comprehensive understanding of this essential human ability.