John Mikhail on Moral Intuitions

Florian Demont
(University of Zurich)
floriandemont232@gmail.com

John Mikhail's *Elements of Moral Cognition. Rawls' Linguistic Analogy and the Cognitive Science of Moral and Legal Judgement* is an ambitious book. It combines themes from Rawls' moral philosophy with cognitive science along Chomskian lines. Instead of bickering about Mikhail's reading of Rawls or the Chomskian framework in general, I start by assuming that Mikhail's exegeses are correct and that the Chomskian framework can in principle be used to study moral and legal intuition.[1] Based on this, I shall draw attention to a specific requirement that Mikhail's theory of moral cognition is supposed to fulfil. The requirement is that the cognitive processes constituting moral and legal intuitions must generate deontic structures (understood to be oughts as manifested in the intuitions). I then go on to argue that Mikhail has not shown that his account of the relevant cognitive processes does generate such structures. The primary upshot of the paper will be that Mikhail must substantially revise his account of how oughts enter moral and legal intuitions. I shall also raise three other objections, which question the philosophical underpinnings of Mikhail's project.

Let us start with an outline of Mikhail's project and its philosophical underpinnings. The primary aim of Mikhail's monograph is to propose and motivate a specific research project which combines elements from moral philosophy, cognitive sciences and legal theory. Mikhail argues at length that the most important philosophical insights underlying the project can already be gathered from the early works of John Rawls. One might want to object to

[1] Mikhail does, as far as I can see, not distinguish between the terms 'judgement', 'intuition' and 'sentiment'. I shall prefer the term 'intuition'. Whatever else one might mean by the term, I shall use 'intuition' here to refer to dispositions to make moral or legal judgements.

this exegetical claim, but such an objection is not what I am after here. There are two reasons why I neglect possible questions concerning Mikhail's Rawls exegesis. First, Mikhail's project would still be interesting if it was not inspired by Rawls. And second, the way Mikhail brings cognitive science to bear on moral theory is a distinct feature of his own and it is exactly this aspect of Mikhail's project that interests me most.

Mikhail's research project is to give an account along Chomskian lines of moral and legal intuitions. Noam Chomsky has argued that a formal grammar can describe and explain the most central aspects of human language. Such a formal grammar is thought to achieve its goals by tracking down the unconscious processes responsible for those linguistic dispositions, which form the kernel of human language.[2] In analogy to this line of thought, Mikhail seeks to describe and explain moral and legal intuitions by giving a formal account of the unconscious processes, which generate them. He sometimes calls this idea the 'moral grammar hypothesis'.

An important feature of Chomskian linguistics is that it is not only naturalist – i.e. it does not only treat the linguistic properties it examines as, for example, physics treats the properties it examines – Chomskian linguistics is also internalist: linguistic properties are to be explained in terms of internal states of organisms.[3] And internal states are computational states as described by a formal grammar.

The moral theory that Mikhail is after is also internalist and he calls it a 'theory of I-morality'. Mikhail hence seeks to explain moral and legal intuitions in terms of internal states of organisms. And again, the relevant states are computational states. But unlike Chomsky, Mikhail does not spend any time defending the conceptual soundness of the Chomskian framework. This is understandable, because defending internalism against philosophical nagging is not an easy task.[4] But it is also highly problematic, because philosophers have produced different powerful objections against Chomskian approaches to human linguistic and psychological capacities.[5]

[2] The locus classicus of this claim is Chomsky, 1965. An accessible up-to-date account of the Chomskian perspective is Hauser et al., 2002.
[3] Chomsky, 2000, 134.
[4] It has been argued that it is in fact an impossible task as long as internalism requires ascribing such properties to brains which can only be ascribed to persons (cf. Demont, 2012). But objections along these lines will simply be bracketed for what follows below.
[5] Some of the objections Mikhail should have discussed are those raised in Hacker, 1990, Smith, 2006 and Wright, 2001.

In what follows below I shall simply assume that the Chomskian framework does provide good descriptions and explanations of some linguistic and mental properties. I also grant to Mikhail that it is conceivable that the framework might provide good descriptions and explanations of at least some aspects of moral and legal intuitions. I make these assumptions for the sake of exposition, as it helps bringing to the fore what I regard as the most problematic aspect of Mikhail's theory of I-morality without getting tangled up in familiar problems he shares with most other Chomskians. By this I do not want to insinuate that Mikhail may neglect these problems, I merely claim that they are somewhat independent of the objections I want to raise here.

There are two constraints on a theory of I-morality, which, if they are met, make the theory empirically adequate. The first constraint is descriptive adequacy and the second constraint is explanatory adequacy.[6] A theory of I-morality is descriptively adequate if it correctly describes human moral and legal intuitions. It is, however, not enough to provide a listiform account, which merely enlists data about moral and legal intuitions. Mikhail, following Rawls and Chomsky, is after a set of principles, which mechanically enumerate all possible moral and legal intuitions and, at the same time, assigns a structural description to each possible intuition. These principles allegedly transcend what people are aware of in everyday life – and this is thought to constitute a first reason why common sense is of limited import in assessing a theory of I-morality.[7]

A theory of I-morality is explanatorily adequate if it provides a description (in the sense just introduced) of the initial state of a human being's moral and legal sense and if it can explain how actual moral and legal intuitions are extrapolated from that when a child grows up and learns to interact with its environment. Meeting explanatory adequacy requires a considerable amount of idealisation and theory-forging by moral psychologists, because the initial state is simply not something which can be directly observed.[8] The idealisations of moral psychologists are also thought to move the subject

[6] Mikhail is aware that empirical adequacy might also require a story about how moral and legal intuitions have evolved in the species and how such intuitions are physically realised in the brain (p. 29 and fn.12 on p. 30). This sort of simplification is quite common within Chomskian circles and it is arguably a good idea to study cognitive processes somewhat independently of biological processes (cf. Demont, 2012).
[7] Mikhail, 2011, 48-51.
[8] Mikhail, 2011, 22.

matter of the theory further away from what can be assessed in terms of common sense.

Together, descriptive and explanatory adequacy yield empirical adequacy. Empirical adequacy is, of course, not enough to make Mikhail's theory of I-morality a *moral* theory. A moral theory must at least be able to explain what counts as a justified moral principle. Mikhail calls this requirement 'normative adequacy'. A more powerful moral theory also gives an account of how moral principles can be justified. A moral theory meeting this constraint is metaethically adequate.

There is an important *prima facie* problem concerning metaethical adequacy that arises if both empirical and normative adequacy are to be met. An empirically adequate account explains moral intuitions in terms of what is. Normative adequacy, on the other hand, requires an account of what ought to be, simply because rational creatures ought to do whatever is in accord with justified moral principles. If such oughts are to be derived from an empirically adequate moral theory, then Mikhail must say something about why that is possible. He must, in other words, solve the is-ought problem. The problem is a real one for him, because Mikhail explicitly claims that the 'descriptive takes precedence over the normative'.[9] After all, it has been a common place for a long time that deriving oughts from what is (or from descriptions) requires a substantial amount of philosophical argument in its support. Applying Moore's take on the issue, deriving oughts from what is (or from descriptions of what is) is not possible, because a computational story about human moral and legal intuitions may settle what our intuitions are, but we can then still question whether these intuitions are good or bad. And Mikhail must show that, at least in some paradigmatic cases, such an open question makes no sense.[10] So, how does Mikhail respond to this?

A theory of I-morality is metaethically adequate if it explains how moral principles can be justified. The problem is that a theory of I-morality must at the same time be descriptively and normatively adequate. In order to solve the problem, Mikhail takes his cues from Rawls:

> As I interpret him, Rawls presupposes a complicated answer to the general problem of justifying moral principles, which turns on at least three potentially unrelated ideas: first, that moral principles can be presumptively justified by showing that they are a solution to the problem of explanatory adequacy; second, that descriptively adequate moral principles can be further justified by showing that they are part of a solution to the problem of explanatory adequacy;

[9] Mikhail, 2011, 30.
[10] Compare Miller, 2004, 13-15 for an introduction to Moore's open-question argument.

> and third, that moral principles that meet the demands of descriptive and explanatory adequacy can be justified to an even greater extent by showing that the adoption of such principles can be proven as a formal theorem in the theory of rational choice. As I understand it, Rawls' notion of *reflective equilibrium* is intended to suggest that these three apparently disparate ideas can, in fact, be reconciled. In other words, Rawls assumes as a general matter that the same set of moral principles can be part of a single, comprehensive solution to the problems of descriptive, explanatory, and normative adequacy simultaneously.[11]

An important idea here is that justifying some moral principles is provisional or presumptive. If open questions about the justification of a moral principle are always possible, we merely need a metaethics, which is good enough for whatever our purposes are. And in Mikhail's scheme of things, (descriptively adequate) moral principles are justified enough if they are 'part of a solution to the problem of explanatory adequacy'. Now, scientific theories like the theory of I-morality are always open to revision, but that does not mean that such theories are not practically applicable, because open questions remain about the justification of the theory's moral principles. The theory of I-morality fulfils its purposes as long as it contains those moral principles, which 'free and equal' persons would regard as a rational choice.[12] Mikhail does not spell out in detail what 'free and equal' means here, but he suggests that the pieces fall into place once we understand the philosophical ideal of a reflective equilibrium.

So, how does Mikhail construe the notion of a reflective equilibrium, how does this notion relate to the theory of I-morality and does it solve the problem of metaethical adequacy? Here are two representative quotes from the monograph:

> [R]eflective equilibrium is a *technical* concept in Rawls's framework, which strictly speaking refers to a state of affairs rather than a method or technique: namely, the state of affairs in which moral principles and considered judgements coincide, and the researcher thus understands the principles to which those judgements conform, together with the premises of those principles' derivation (Rawls 1971:20). Moreover, Rawls defines the meaning of reflective equilibrium in the context of a conception of moral theory whose principal aim is to solve the problems of empirical and normative adequacy with respect to I-morality.[13]

> [T]he primary function of the concept of a considered judgement in Rawls' framework is to select, from among the moral judgements people *actually* make,

[11] Mikhail, 2011, 31.
[12] Mikhail, 2011, 32.
[13] Mikhail, 2011, 289.

> those judgements that the moral theorist believes are truly evidential, insofar as they reflect the properties of an underlying cognitive competence.[14]

From this we can gather two claims. First, a reflective equilibrium is the state of affairs in which moral principles and considered judgements coincide. But they only should be said to coincide if the moral psychologist thereby understands the computational processes generating the judgements and the elements from which the judgement is generated. The second claim is that considered judgements are those moral and legal judgements, which people actually make and which a moral psychologist takes to be evidential. If we sum this up, the reflective equilibrium turns out to be the state of affairs in which a theory of I-morality generates exactly those moral principles, which account for the moral and legal intuitions that moral psychologists have regarded as relevant data.

This solution to the problem of metaethical adequacy has three shortcomings. First, consider that the psychologists selecting the data are also the psychologists constructing the theory. The problem with this is that some data count as relevant if the psychologists believe that these data reflect 'the properties of an underlying cognitive competence' to recognise what is morally or legally correct. The psychologists' beliefs about what are relevant data will change as they go along constructing the theory of I-morality, which determines what the properties of the underlying competence are. Mikhail has to make clear that the idealisations governing data selection are appropriately independent of the details of the theory of I-morality that is derived from it. Otherwise, the theory's predictions will be trivially true, as only those findings count as relevant data which theory predicts. This is a general problem about idealisations in scientific reasoning and it is surprising that Mikhail has not addressed it in the monograph. It is hard to see how one can argue that the theory's predictions are not trivially true if neither data gathering nor theory construction is sensitive to common sense conceptions of what is relevant regarding moral an legal intuitions.

A second shortcoming of Mikhail's solution to the problem of metaethical adequacy has to do with the concept of computability. It is a formal truth about computational processes that we cannot mechanically determine in advance whether a specific mechanical derivation will eventually produce a definite solution or not.[15] Applied to I-morality, it might be that a computational account

[14] Mikhail, 2011, 283.
[15] This is, of course, the halting problem (cf. Boolos, Burgess & Jeffrey, 2007, 40).

can be given of how a specific moral or legal intuition is generated, but we will not be able to determine through any (Turing-computable) derivation whether or when the intuition will effectively be generated. Yet a rational person can be expected to always have a definite moral or legal intuition when presented with a case, of which she understands all elements and where she has to make a judgement. If we present a rational person with a formal description of a case, she will be able to tell whether she can make a moral judgement if the details are filled in, but we can make no such prediction about computational derivations of moral and legal judgements. So it appears that a rational person always knows more about her moral and legal intuitions than a theory of I-morality can capture. Now, recall that a theory of I-morality is descriptively adequate only if it correctly describes human moral and legal intuitions. With this requirement in the background it appears that descriptive adequacy can never be achieved by a theory of I-morality, because a rational person's knowledge of her moral and legal intuitions is necessarily richer than what a computational account can capture.

The third shortcoming harks back to the is-ought problem outlined at the beginning. We may ask whether Mikhail's conception of a reflective equilibrium tells us how any oughts can be derived from what is. And the answer to this is: no. In a reflective equilibrium, moral principles and considered judgements simply coincide. Any oughts playing a role must have been in the moral principles and considered judgements before they were compared. According to Mikhail's conception of I-morality, deontic structures – moral and legal oughts as they figure in moral principles and considered judgements – are generated whenever moral and legal intuitions are generated. So it is not the conception of reflective equilibrium which tells us how to solve the problem of metaethical adequacy, but the details of the derivation process. The big question now is whether the derivation process can solve the is-ought problem.

In chapter 6.5, Mikhail gives a very short and abstract account of how moral and legal intuitions are derived. If a rational person perceives a situation calling for a moral or legal judgement, her cognitive processes must first identify relevant descriptions of the action. With these descriptions, the cognitive processes must temporally order particular events. On the basis of that, the cognitive processes must identify causal structures in the temporally ordered events. The next step is a distinct element of a theory of I-morality: moral structures are identified by labelling certain effects as good or bad. So, if somebody dies as an effect of some other event, that effect will be regarded

as bad. If, on the other hand, somebody is saved from death as an effect of some other event, that effect will be regarded as good. Note that this does not yet generate oughts. Simply labelling some effects within a causal structure does not suffice to establish an ought, because it is not made clear by this what should be done. That homicide, for example, is a bad effect in any series of causally related events may still be part of the representation of what is the case.

In order to derive oughts, Mikhail first introduces intentional structure. The derivation of intentional structure is accounted for under a presumption of innocence or good intention. Every bad effect will be labelled as a side effect of some actions and every good effect will be labelled as an end or goal. Such a presumption of innocence is, of course, highly problematic, because pursuing good and avoiding evil will be a matter of 'innate instinct'.[16] Somebody who seeks to do evil will, according to Mikhail's proposal, turn out to have defective instincts. The right way of dealing with such wrongdoers is to give them treatment and not to punish them. This is, however, not at all how matters are handled. We distinguish between moral defects and cognitive defects, because any wrongdoing will be punished less severely if the agent has a cognitive defect. People with moral defects will, however, be punished more severely to keep them from harming themselves and others in the future. Mikhail's proposal does hence play down a perfectly sensible distinction between, on the one hand, what is morally and legally wrong and, on the other hand, what is a cognitive defect. They may be related in some cases, but they must not be run together as a matter of principle.

Deontic structure is now explained based on this problematic notion of intentional structure. The point of the last step of the derivation process is to come up with some sort of explanation of intuitions about what is permissible, forbidden or obligatory. Mikhail helps himself to some basic legal definitions to achieve this:

> One key insight of the moral grammar hypothesis is that adequate structural descriptions must also incorporate prima facie legal wrongs, such as battery or homicide.[17]

Assuming for a moment that all the structures (including intentional structures) can be had just as Mikhail thinks, how do we incorporate prima facie legal wrongs, such as battery, within the confines of I-morality? We

[16] Mikhail, 2011, 173.
[17] Mikhail, 2011, 173.

presumably have available representations of events such as '*X* throws *Y* off a foot bridge', '*Y* prevents a train from hitting 5 innocent people' and '*X* kills *Y*'. Now '*X* throws *Y* off the footbridge' (event 1) came before and caused '*X* kills *Y*' (event 2). Event 2 will be labelled a bad effect of event 2. But event 2 will count as a side effect of event 1, because event 1 also caused the good event '*Y* prevents a train from hitting 5 innocent people' (events 3). Event 3 will then count as a good effect of event 1 and it will thus count as an end. Adducing prima facie legal wrongs, we can now explain what counts as forbidden and then, in a later step, derive permissions (i.e. that which is not forbidden) and obligations (i.e. that which it is forbidden not to do).

Now, *X* ought not to throw *Y* off the footbridge, because he thereby commits battery.[18] The legal definition of battery, which Mikhail adduces to derive the ought, has it that *X* commits battery if he touches *Y* without his 'express, implied or hypothetical consent'.[19] But how does a rational person come to know prima facie legal wrongs such as battery? After all, the theory of I-morality is internalist – it explains moral and legal intuitions in terms of internal states of organisms – and it is not clear how legal definitions, like the definition of battery, are related to internal (computational) states of organisms. I can see two possible replies that Mikhail can make. First, he could claim that the concept of battery is innate. Second, he could concede that the concept of battery has been internalised in some way.

So, what if the concept of battery is innate? The problem with this first possible reply is that Mikhail uses official legal definitions in his derivations. It is quite hard to imagine what would count as verifying that a child has a tacit grasp of battery as touching somebody without that person's express, implied or hypothetical consent. If one insists that the concept is innate and that the legal definition was merely one way of expressing that concept, we must ask whether any empirical finding could topple that claim. There is no such empirical finding, because any alleged evidence against nativism about legal concepts can be rejected by insisting that the data has not been selected in accord with the sort of idealisations that I-morality requires. The account of deontic structures would then be unfalsifiable and a theory of I-morality cannot be, at the same time, empirically adequate and feature an unfalsifiable

[18] Of course, X may throw Y off the footbridge if committing battery and killing him is a side effect of stopping a train, which would otherwise kill 5 people. This more complicated case does, however, not help elucidating the aspect of the derivation of deontic structures that interests me here.
[19] Mikhail, 2011, figure 6.2g.

account of deontic structures as an alleged *result* of empirical enquiry. So, nativism about basic legal wrongs does not have good prospects.

And what if the concept of battery is not innate, but has been internalised in some way? The problem with this second possible reply is that Mikhail's theory of I-morality is an internalist theory. If the derivation of deontic structures requires adducing definitions of prima facie legal wrongs and if these definitions have to be learned (by reading, for example, the Second Restatement of Torts and, perhaps, by looking at court cases where the definitions are applied), then the derivation of deontic structures cannot be explained in purely internalist terms. This amounts to rejecting unbridled internalism for a viable theory of I-morality.

To sum all of this up, we can make four different objections to Mikhail's proposal. The most important one is that he has not solved the is-ought problem. If the derivation of deontic structures builds on innate legal concepts, the theory of I-morality is not empirically adequate, because the derivation bridging the is-ought gap is part of an unfalsifiable empirical claim. On the other hand, if legal concepts are acquired through some sort of training, a correct moral theory cannot be purely internalist.

The second objection was that the derivation of moral and legal intuitions should not blur the distinction between moral and legal wrongs on the one side and cognitive defects on the other side. An interesting proposal such as Mikhail's must not provide a possible excuse for people who neglect their moral responsibilities and who want to lower their liability by pointing out that they are victims of defective instincts. Whatever moral theory one proposes, one should always make sure that it remains applicable.

The third objection was that a computational derivation of moral intuitions is subject to formal constraints on computability – especially the halting problem – whereas a rational person's knowledge of her moral and legal intuitions is not subject to these constraints.

The fourth objection was that Mikhail has not made clear that the process of dividing available data into relevant and irrelevant data is sufficiently independent of a theory of I-morality so as to ensure that the predictions of the theory are not trivially true. It might turn out that this objection cannot be met unless idealisations are made sensitive to what is acceptable from a common sense point of view.

I do think that Mikhail's monograph contains good general ideas and that the whole project is exemplary in how it brings together insights from

philosophy, psychology, linguistics and legal theory. But I also think that there are substantial problems when it comes to the details of Mikhail's theory of I-morality. It might be possible to explain away some problems and it might also be possible that other problems can be met through further research, but Mikhail does have to rethink the conceptual basis of his project.

References

Boolos, G.S., Burgess, J.P. & Jeffrey, R.C., 2007, *Computability and Logic. Fifth Edition,* Cambridge, Cambridge University Press.

Chomsky, N., 1965, *Aspects of the Theory of Syntax,* Cambridge (Mass.), MIT Press.

Chomsky, N., 2000, *New Horizons in the Study of Language and* Mind, Cambridge (Mass.), MIT Press.

Chomsky N. & McGilvray J., 2012, *The Science of Language,* Cambridge, Cambridge University Press.

Demont, F., 2012, Chomsky's Methodological Naturalism and the Mereological Fallacy. In: *Philosophical and Formal Approaches to Linguistic Analysis*, Frankfurt a.M., Ontos.

Hacker, P.M.S., 1990, Chomsky's Problems. *Language & Communication* (10)2, 127-148.

Hauser, M.D., Chomsky, N., & Fitch, W.T., 2002, The Faculty of Language: What it is, Who has it, and How did it Evolve? *Science* 298, 1569-79.

Lepore, E. & Smith, B.C. (eds.), 2006, *The Oxford Handbook of Philosophy of Language.* Oxford, Clarendon Press.

Mikhail, J., 2011, *Elements of Moral Cognition. Rawls' Linguistic Analogy and the Cognitive Science of Moral and Legal Judgement.* New York, Cambridge University Press.

Miller, A., 2004, *An Introduction to Contemporary Metaethics.* Cambridge, Polity Press.

Rawls, J., 1971, *A Theory of Justice*. Cambridge (Mass.), Harvard University Press.

Smith, B.C., 2006, What I Know When I Know a Language. In: Lepore and Smith (eds.) 2006, 941-82.

Wright, C., 2001, *Rails to Infinity*. Cambridge (Mass.), Harvard University Press