Please cite the published version, which appeared in Philosophy of the Social Sciences.

The Incentivized Action View of Institutional Facts and the Searlean View: A Response to Butchard and D'Amico

JP Smit, Filip Buekens and Stan du Plessis

Abstract

In Smit et.al. (2011, 2014) we argued, *contra* Searle, that institutional facts can be understood in terms of non-institutional facts about actions and incentives. Butchard and D'Amico (2015) claim that we have misinterpreted Searle, that our main argument against him ('the circularity objection') has no merit and that our positive view cannot account for institutional facts created *via* joint action. We deny all three charges.

In Smit et. al. (2011, 2014) we argued against the Searlean view of institutional facts and in favour of what we call the 'incentivized action' view. On this view all institutional facts should be understood in terms of the incentivization of action. We also claimed, *contra* Searle, that such actions can be characterised in terms that are not irreducibly institutional. In their 'Alone Together: Why "Incentivization" Fails as an Account of Institutional Facts' (2015), Butchard and D'Amico take issue with our views. They claim that we have misunderstood Searle's claim that institutional reality constitutes a 'huge, invisible ontology' and deny that Searle's claim that institutional terminology cannot be characterized in non-institutional terms raises the problem of definitional circularity. They also claim that our view cannot account for joint action. We are, for the most part, unmoved.

1. Searle's 'huge, invisible ontology'

Searle has repeatedly characterized institutional facts as constituting an 'invisible ontology' (1995: 3-4; 2005: 1). Butchard and D'Amico charge with us with having misconstrued this claim as indicating that Searle's adopts a non-naturalistic ontology (317). They interpret

Searle as saying that it is not the case that a screwdriver *qua* social object is a distinct thing from a screwdriver *qua* physical object. Searle thinks that there is just one object, the screwdriver, that possesses both the property of being a molecular structure and the property of being an artefact (317).

The charge as stated is inaccurate, though a modified formulation has some merit. We did not attribute to Searle the claim that when dealing with traffic lights, money and borders, in each case we somehow have two ontologically distinct objects. We interpret Searle as saying that, in such cases, there is only ever one thing, *though individuated by distinct properties*. We are surprised that the authors accuse us of such a misinterpretation as we have been very explicit on this issue. Butchard and D'Amico quote pages 3 - 4 from our 2011 paper in support of their claim; two pages later we dedicate a numbered section to exactly this issue. We say:

> [Searle claims that] talk about social objects is just talk about social facts, and the descriptions describing these facts ultimately function only as a different way of potentially picking out the same natural object. In this regard, we fully agree with Searle. (2011: 6).

We explicitly state that we agree with Searle that, in the case of something like a traffic light, we are not dealing with two distinct objects. Rather we have a single object which is an institutional object in virtue of some non-intrinsic property of it (2011: 6 - 7). We interpret Searle as saying that this property is a matter of being the object of a collective attitude that can only be expressed using irreducibly institutional concepts. Our alternative theory is that this property is a matter of being an object that people have been incentivized to act towards in some relevant way (2011: 6 – 7)[1].

While we did not commit the error of interpreting Searle as saying that a traffic light is somehow two things, we are guilty of a more subtle error. Searle frequently says that, in the case of institutional reality, 'the fact can only exist as far as it is represented as existing' (2005, 13), i.e. we make something the case by representing it as being the case.

[1] Such concepts reflect a complex profile of incentivized actions associated with an object, place, or person. A similar idea is developed and defended in Guala (2014) and Guala and Hindriks (2014).

This claim can be interpreted in two ways. Consider money; on the first interpretation 'being money' is fully reducible to, i.e. identical with, 'being collectively represented as being money'. Call this the *deflationary reading*. On the second interpretation 'being collectively represented as being money' may be necessary and, given suitable background conditions, sufficient for 'being money', yet not reducible to it or identical with it. Call this the *strong reading*. On such a reading Searle is committed to the existence of irreducibly institutional facts, whereas on the deflationary reading he is committed to merely the existence of irreducibly institutional concepts. Note, of course, that this distinction does not concern objects, but properties, i.e. neither reading claims that a traffic light is somehow two things.

We have not always clearly distinguished the above two interpretations. In passages like the one discussed above we adopted the deflationary reading, but have frequently expressed our claims in terms of 'institutional facts', not 'institutional concepts'. Note, however, that this does not affect our main argument against Searle, namely that his positions lands him in a definitional circle. This charge applies equally to both readings. Our point about parsimony is similarly unaffected; whether our position rids one of irreducibly institutional facts or irreducibly institutional concepts matters little. There is, however, a genuine problem here: which reading is the correct interpretation of Searle?

The deflationary reading is the more naturalistic one. On this view institutional reality boils down to no more than brute objects and collective attitudes, with the proviso that the content of such attitudes cannot be stated using non-institutional concepts. On the strong reading matters are far more mysterious as human beings then literally have the ability to create facts merely by representing them and yet such facts are ontologically distinct from their representations. The deflationary reading is the interpretation of Searle we outlined in the passages cited from our 2011 paper. One would think, based on his general commitment to naturalism, that Searle would endorse it. We cannot, however, claim this with any great confidence. He could also endorse the strong reading, but deny that doing so violates his naturalistic scruples. The situation is interestingly analogous to his view on consciousness, which features a famously accommodating version of naturalism in which mental facts are irreducible to other kinds of facts, yet this is claimed to present

no great difficulty.[2]

Can Searle consistently endorse the deflationary reading? It is not clear that he can. Searle characterizes his naturalism in terms of a commitment to physicalism[3]. Yet sometimes he explains his view of institutional reality by saying that the creation of institutional objects should be understood as a matter of creating deontic powers, i.e. rights, duties and obligations (2005: 13). Such deontic powers are claimed to be 'at the heart of the institutional reality' and are explicitly characterized as 'abstract entities' (Smith and Searle, 2003: 305). It is not clear how such abstract entities are supposed to fit into his naturalism or how one can claim this and still endorse the deflationary reading[4].

More worrying is Searle's distinction between observer-relative and observer-independent facts. He writes:

> A rough test for whether or not a phenomenon is observer independent or observer relative is: could the phenomenon have existed if there had never been any conscious human beings with any intentional states? On this test, tectonic plates, gravitational attraction, and the solar system are observer independent and money, property, and government are observer relative. The test is only rough-and-ready, because, of course, *the consciousness and intentionality that serve to create observer relative phenomena are themselves observer independent phenomena.* For example, the fact that a certain object is money is observer relative; money is created as such by the attitudes of observers and participants in the institution of money. *But those attitudes are not themselves observer relative*; they are observer independent (2005: 3-4, our italics).

The above passage seems unambiguous as can be. Searle thinks that conscious, intentional states are observer-independent facts, yet the fact that a certain thing is money is observer

[2] See Rust (2011) for a discussion of this distinctively Searlean strategy.
[3] "How can we accommodate a certain conception we have of ourselves as conscious, mindful, rational, speech act performing, social, political, economic, ethical, and free-will possessing animals in a universe constructed entirely of these mindless physical phenomena?" (Searle, 2005: 5)
[4] Searle makes this claim when pushed on matters concerning the existential commitments of his view by Barry Smith (Smith and Searle, 2003). Also see Smith (2008) for an in-depth discussion of whether Searle's ontology is coherent.

relative. Or, to use our earlier formulation, 'being money' is an observer-dependent property, yet 'being collectively regarded as being money' is an observer-independent property. But then the deflationary reading, despite its evident attraction, is off the table; the two properties identified with one another by the deflationary reading, namely 'being money' and 'being regarded as money', cannot be identical as *they are not even of the same ontological kind*.

Based on the above it looks like Searle's wishes to adopt the strong reading. His general naturalism, however, suggests that he would prefer the deflationary reading. We do not know which reading is correct, or if he is simply inconsistent. The more basic question, however, is whether either reading provides a viable theory. This issue brings us to our critics' second objection.

2. Definitional circularity

In Smit et. al. (2014) we explained that part of our reason for abandoning Searle's account is that, on his view, institutional reality can only be characterised by using irreducibly institutional concepts. This immediately raises a worry of definitional circularity; if I can only explain what money is by saying that it is something collectively represented as being money, then the concept money has not been usefully clarified. Searle recognized this issue (1995: 52 – 53) and stated that it is not problematic as money can be defined in terms like 'buying', 'selling', 'owning', etc. We do not think that this helps and so abandon his view altogether.

Butchard and D'Amico object to this by saying that Searle does not think that explaining 'money' in terms of 'buying', 'selling', etc. results in a definitional circle, even if this is unproblematic. Rather he says that the term 'money' marks a node in a network of practices like buying, selling, and so on. Furthermore, there is no reason to think that some nodes cannot be grasped independently of others and so no issue of circularity arises (320).

We do not think that their reasoning is completely persuasive. Given that Searle states

that explaining money in terms of buying, selling, etc. is a matter of 'expanding the circle' (1995: 52), we took him to acknowledge the existence of a circle, but denying that it is problematic. Butchard and D'Amico admit that he refers to this as 'expanding the circle', but deny that he should be interpreted as acknowledging circularity, even of an innocent sort. Their argument is that Searle says that one can grasp one node independently of grasping the others (320). On this interpretive issue they may well be correct. It is at least possible that Searle's use of 'expanding the circle' is misleading and that he does not think that there is circularity involved here, whether vicious or otherwise. Even if this is the case, however, we stand by our charge of circularity. The crucial issue is whether it is really possible to grasp one node independently of grasping the others. If so, then the charge of circularity collapses as nodes that cannot be independently grasped can then be explained in terms of nodes that can be so grasped. But we simply do not see how this is supposed to be possible.

We will first deal with this issue as it arises within the deflationary reading. On this view something being money is reducible to the fact that we collectively represent it as money. What, however, is the content of this propositional attitude supposed to be, i.e. what are we representing it *as* if we represent something as money? All parties agree that the 'moneyness' of some object used as money (a shell, a cigarette, a banknote) is not some ordinary physical property of it. Furthermore, on the deflationary reading, there is no extra-conceptual reality above and beyond the object, its properties and the relevant propositional attitudes. This raises a basic problem for Searle. On such a reading, institutional objects are no more than brute objects individuated by being the attitude of a collective attitude featuring irreducibly institutional concepts. Hence the term 'money' cannot be fully defined in terms of its extension, i.e. what the property 'being money' applies to in reality. The irreducibly institutional part of the concept 'money' does not latch on to anything outside itself, i.e. it has a *mind-to-nothing* direction of fit. Saying that money is a node in a network of practices does not help. The problem repeats for any other node (buying, selling, etc.) that we care to mention as *the exact same thing is true of it*. To the degree that the terminological web which is supposed to explain 'money' is irreducibly institutional, there is, on the deflationary reading, *nothing* in any actual practice that we *could* grasp in order to fully explain the concepts to which the

explanatory buck is passed. Hence trying to alleviate the problem with reference to some other node cannot do anything useful. We will only get stuck in a circle. Or, alternatively, in an infinite regress.

Butchard and D'Amico claim that, by our standards, our view is also circular (321). We rely on notions like incentivization and, as they say, such terms can only be understood in terms of other psychological notions like desires beliefs, and so on. Does this open us to the charge of circularity? No, definitional links only provide a *prima facie* problem if there is nothing outside them to explain how the terms in question have their contents. We have no reason to think that psychological terminology has this problem, or can only be reduced to terminology that has this problem. While we have no firm stance concerning the ontological status of beliefs and desires, our view is compatible with all views that we are aware of. Unless Butchard and D'Amico can show that *all* views open to us have the consequence that psychological notions are irreducibly psychological, yet don't latch on to anything real in the world, the issue simply does not arise. Further, note that philosophers are committed to giving an account of beliefs and desires in virtue of non-institutional reality. So, *even if* their charge had merit, our reduction would still serve to show that a seemingly new problem actually reduces to an old problem.

One way out of the problem concerning the 'mind-to-nothing' direction of fit would be to claim that the conceptual links between nodes is itself fully constitutive of the content of such terms. But then such terms are purely formal devices, i.e. symbols and strings that serve to transform other symbols and strings. We do not know of any theory that claims that the irreducibly institutional nature of institutional reality can be captured in this manner, or see how an argument for such a view could be developed.

At the very least, the Searlean theorist owes us an account of how the irreducibly institutional concepts can have any content. Maybe the theorist could try and explain how such terms have their contents by adopting some sort of fictionalism about institutional objects. Or, perhaps, the theorist could liken institutional terminology to taking an 'institutional stance' (similar to Dennett's 'intentional stance') to a social practice.

Another option would be to define the relevant concepts in terms of their functional role[5]. In any case, there is a large explanatory burden here that has not been discharged. Simply passing the buck to other terms that raise the exact same issue is of no help. Our charge of circularity stands.

In summary, Butchard and D'Amico claim that some nodes in an institutional network can be grasped independently of other nodes. Our problem with this is that, on the deflationary reading, which they support[6], and to the degree that such concepts are irreducibly institutional, there is nothing to grasp.

An alternative way to avoid this problem would be to adopt the strong reading. On this view institutional facts have an ontological status over and above that of the relevant brute objects and propositional attitudes. Note, however, that on such a reading our view *does* constitute a genuine ontological reduction and, if feasible, is preferable on grounds of parsimony. Moreover it is not clear that such the strong reading would really help resolving the semantic issue explained above. For I would have to grasp the nature of money prior to representing it as existing, yet the term 'money' would only have meaning subsequent to being so represented[7].

Both the strong reading and the deflationary reading result in deep difficulties concerning definitional circularity or infinite regress. Butchard and D'Amico's claim that we can grasp some nodes independently of others is a mere assertion, backed by no argument. This is one of the reasons why we jettison the idea of an irreducibly institutional terminology.

3. Joint action and institutional reality

Butchard and D'Amico also object to our positive view. They base their argument on the nature of joint action. A 'joint action' is a matter of doing things together, i.e. taking a walk together, and so on. They believe that such actions cannot be understood if we do

[5] Then, however, the exact same problem would recur once we try to characterize the relevant functions.
[6] They correctly point out that on such construal, Searle's view is itself reductive as institutional facts are reduced to brute objects and propositional attitudes (321).
[7] The same problem holds even if we only require that the experts in a linguistic community grasp the relevant concept.

not endow the parties to the joint action with a joint intention. Furthermore, they believe that such joint intentions cannot be fully analyzed in terms of individual intentions (233 - 324). They further believe, though they need not endorse for their argument, the view that this distinction between joint and individual intentions can be captured by a distinction between intention of the form 'we intend..' and 'I intend...' (322). They object to our characterization of institutional reality in terms of actions and incentives by claiming that our analysis cannot account for such joint intentions and joint actions.

We do not think that their argument succeeds. Nothing in our view commits us to any specific treatment of joint intentions, joint actions and so on. In fact, the charge relies on a fundamental misunderstanding of our theoretical aims. A reminder of the dialectic thus far: Searle claims that, first, all institutional reality is created in virtue of collective intentionality and, second, that it can only be captured in irreducibly institutional terms. Our project has mostly been motivated by the latter claim and it is this claim that we take ourselves to have refuted. We also challenge the first claim. This, however, is *not due to our views on the nature of collective intentionality*. We have, throughout, pronounced ourselves agnostic (2011: 2 , 2014: 1817) as to the nature of collective intentionality and take no stance on the relation between I-intentions and we-intentions. Our view, rather, is that collective intentions are not *constitutively* required for institutions to exist. This was demonstrated by the 2011 paper in which we discussed the simple example of a traffic light. We showed that at least one instance of one institutional fact can be accounted for without collective intentionality. Given the generality of Searle's claims, i.e. his claim that all of institutional reality is created in virtue of collective intentionality, even one such example is sufficient to refute his view and show that collective intentionality is not a *necessary* ingredient of *all* institutional facts.

We have also made the further claim that we do not see any reason to suppose that *any* institutional fact necessarily involves collective intentionality. This claim has been made good in an essentially inductive way by, when analyzing institutional facts like traffic lights, money, borders, property, promises and companies (2014), always doing so without endowing any of the hypothetical beings involved with such collective intentions. We see no reason to doubt that this is possible; we view the issue of collective intentionality as

orthogonal to the issue of institutional reality. Our motive in carrying out such analyses, however, is not a matter of antipathy to collective intentionality as such. Rather the problem is that collective intentionality has often seemed to be the magic that is supposed to make Searle's irreducible institutional facts/concepts somehow less strange. We try to loosen the connection between collective intentionality and irreducibly institutional facts/concepts in order to further discredit the latter, not to undermine the former.

The above is reflected in our positive view, which is that institutions can be characterized in terms of actions and incentives. Our positive view as such contains no positive assertions about intentionality, whether collective or otherwise. Even if we are wrong in our view that all institutions can, in principle, arise in virtue of only individual intentions, our positive view would remain *entirely* unchanged. It would just turn out that some of the incentivization happens in terms of collective intentions. In fact, even if it turns out that all institutions require collective intentionality, our positive view would survive unscathed.

Butchard and D'Amico state that we 'reduce the collective action to lower-level intentions that... do not require we-intentions' (323). This, as explained above, is not the case as we never try to reduce collective intentionality to anything else. Rather we try to show that, for whichever institution is under discussion, that it *could* have come into existence without 'collective action' (in their 'strict' sense that definitionally requires collective intentionality) and *hence the matter of collective intentionality is orthogonal to the nature of the institution* under discussion.

They go on to claim that our reduction does not work as we cannot account for collective action. To prevent confusion, note that they are not challenging the reduction of *institutional* concepts to concepts that are not irreducibly *institutional*. 'Walking together' is not an institutional object in the Searlean sense and so we have no reason to try and reduce it to anything else. Searle takes his theory to be about phenomena that fit his *X* counts as *Y* in *C* characterization, i.e. objects (rivers, gold pieces) that we then *count as* being an institutional object (borders, money). Walking together is not such an institutional object. If we intend to walk and then do so together we don't *count as*

walking together, rather we just *are* walking together.

Their objection, rather, is against the attempt to reduce collective intentionality to individual intentionality. They claim that we are unable to do so and that this raises a problem for our notion of incentivization. Of course, as explained above, we did not try to perform any such reduction. Does their argument, however, raise any difficulties for our view?

The argument is that collective action/intention creates duties and obligations, i.e. if we are walking together, then the very 'togetherness' of our walking provides an additional reason to not quit walking. Such duties and obligations constitute reasons for action that are not captured by our notion of incentivization; we 'specify “reasons for acting” too narrowly' (326) for issues concerning collective behaviour to be addressed.

We do not agree. We have always stated that our notion of incentivization is as broad as can be.

> [O]ur talk of incentives reflects the fact that human action is motivated, i.e. based on reasons. These reasons come in many kinds; people may act for reasons that are selfish, altruistic, self-regarding or other-regarding, moral or prudential, and so on. When we say that someone is incentivized to perform an action we merely mean that there is, for that person, some reason for action, *whatever this may be* (2014: 1818-1819, italics added).

We have, furthermore, also been very clear that, on our view, the source of the institutional fact is irrelevant to institutional ontology as such. Our theory concerns ontology and facts about the sources of incentives do not play an individuating role.

> [W]hile the above definitions require that people be somehow incentivized to perform the relevant actions, the source and nature of the incentivization are not individuating facts (2014: 1815).

We don't see how 'walking together' raises any difficulty here. Stipulate that Butchard and D'Amico are correct in their construal of 'walking together' and that it provide a reason for not discontinuing the walk that cannot be captured in any other way. Then we would simply say that the participants to the walk are incentivized to continue walking partly in virtue of the fact that they are walking *together.* We described 'being incentivized to *X*' as meaning no more than 'having a reason to do *X*' (2014: 1818 - 1819). Typically we have formulated the incentives at play in creating institutional reality as being due to 'human agency or moral belief' (2014: 1824). This includes all incentivization that is not completely due to the nature of the intrinsic properties of the institutional object under discussion. The incentivization due to 'walking together' fits seamlessly under the category of 'moral belief'. Any effect that 'walking together' has on actual behaviour can only happen in virtue of the participants' belief, whether implicit or explicit, that 'one should not abandon an act of walking together unless there is an overriding reason to do so'. Such cases are just cases of incentivization *via* moral belief.

The same goes for their later claim that we fail to see that enforcement is not necessary for the creation and existence of a boundary. Searle claims, and Butchard and D'Amico agree, that the joint intention that some lines of stones constitutes a boundary can suffice to make it a boundary. We do not deny any of this. A joint intention can, even in the absence of any conceivable enforcement, give rise to the existence of a boundary. But, again, this would simply be a case of incentivization by moral belief. The boundary will exist inasmuch as the relevant parties believe that 'this line should not be crossed'. This could, in turn, be due to a more general belief that one should act in accord with publicly expressed joint intentions. Such a case would just be another instance of incentivization by moral belief.

Institutions are fundamentally strengthened when people 'buy in' to the institution. Such 'buying in' is, on our view, ultimately a matter of having some relevant moral belief to the effect that the institution is morally virtuous or justified. Such cases are not an objection to our theory, but an integral part of it.

To summarize: Butchard and D'Amico misconstrue our theory as an attempt to reduce

collective intentionality to individual intentionality. Rather we wish to reduce our institutional terminology to non-institutional terminology. This is our positive project; the negative project implicit in it is the view that facts about the nature and source of the incentives that create and maintain institutions are not individuating facts. Such concerns, despite the general impression due to Searle's view, are orthogonal to social ontology. Their construal of collective behaviour does not threaten either our positive or negative project.

Bibliography

Butchard, W., D'Amico, R. 2015. Alone together: why 'incentivization' fails as an account of institutional facts. *Philosophy of the Social Sciences.* 45: 315 - 330.

Guala, F. 2014. On the nature of social kinds. In Galloti, M. and Michael, J. (eds.) 2014. *Perspectives on Social Ontology and Social Cognition*. 57-68. Springer.

Hindriks, F., Guala, F. 2014. Institutions, rules, and equilibria: a unified theory. *Journal of Institutional Economics* (2014): 1-22.

Rust, J. 2011. Review of Franken, D. et. al. (eds.) 2010. "John R. Searle: Thinking about the Real World". *Notre Dame Philosophical Reviews.* Available at https://ndpr.nd.edu/news/24715-john-r-searle-thinking-about-the-real-world/

Searle, J. 1995. *The Construction of Social Reality.* London: Penguin Books.

Searle, J. R. 2005. What is an institution? *Journal of Institutional Economics.* 1: 1 – 22.

Smit, J. P., Buekens, F., & du Plessis, S. (2011). What is money? An alternative to Searle's institutional facts. *Economics and Philosophy* 27, 1–22

Smit, J. P., Buekens, F., & du Plessis, S. (2014). Developing the incentivized action view of institutional reality. *Synthese* 191: 1813 – 1830.

Smith, B. and J. Searle. 2003. The construction of social reality: an exchange. *American Journal of Economics and Sociology* 62: 285–309.

Smith, B. 2008. Searle and de Soto: The new ontology of the social world. In Smith, B. et. al. (eds.). 2008. *The Mystery of Capital and the Construction of Social Reality.* 35 – 51. Chicago: Open Court.