Corporate and Artificial Moral Agency

Alan E. Singer
Holshouser Distinguished Professor of Ethics
Appalachian State University.

Abstract The paper considers the implications of the Corporate Moral Agency debate for the notion of artificial moral agency and the general intelligence project. A distinction is drawn between meta-arguments and object-level arguments, whilst the implications of the arguments within each category are indicated. The “metaphor” “mutuality” and “political” arguments are then discussed further.

1. Introduction

The debate about the application of moral philosophy to non-human entities has been wide ranging and enduring. Within business ethics, it has focused upon the notion of corporate moral agency (CMA) and hence the capabilities, responsibilities, duties, rights and blame-worthiness of corporations per se. Within CIS and cognitive-science, somewhat similar questions have arisen about the notion of artificial moral agency (AMA) as well as the possible moral responsibilities of artificial general intelligences (AGI’s) and their builders. In broadest terms, the present paper considers the relationship between the CMA and AMA debates. However, the primary focus is upon the ways in which the CMA debate (CMAD) informs (i) the very idea of AMA, and (ii) the ethical issues involved in AGI-building (AGIB).

Figure 1. The CMA debate informs AGI/B.

From the outset, it is also worth bearing in mind that if one considers the special case of a virtual firm, “being run by AGI’s” [18, p.313]. In the following section a way of re-structuring the entire CMAD is first proposed

2. Structure

The CMAD is comprised of (i) a set of meta-arguments that are essentially attempts to qualify the importance of the CMA debate itself, and (ii) a set of (object level) pro-CMA and anti-CMA arguments that attempt to establish or else critique the very idea that corporations are moral agents (Figure 2). Various (overlapping) types of argument can then be identified within each set, as listed below and depicted in Figure 2.

Figure 2. The structure of the CMA debate

3. Meta-arguments

There are essentially four main categories of meta-argument. These claim, respectively, that the CMAD is: crucial; a distraction; a mere metaphor, or else misconceived (for other reasons such as “mutuality”).

question around: much has been said about CMA, but what does that tell us about AMA’s?

2 Moore [27] also reviewed some selected pro-CMA and anti-CMA arguments. Any article-length review of the CMAD can cite only a suitable subset of many relevant contributions.

3 Henceforth the acronym AGI/B’s is used, to accommodate the recursive self-building (self-programming) capabilities of AGI’s.

---

1 Hall-Store’s [18] asked “can we say anything about the rights and duties of corporations if...the AI’s will be running them within the next few decades?” The present paper turns that
The first category of argument emphasises the high importance of the entire CMA-debate (CMAD). It is held to be central to the subject of business ethics whilst serving to direct attention towards ethical concerns in business, generally [11]. The CMAD is also seen to be very important in law and practice, due to the need to be able to distribute blame or liability properly across the corporation per se and individual managers, in cases where harms have occurred [14,45]. By implication, a corresponding debate about AMA is “crucial” for similar reasons.

3.1 Distraction

All of the remaining meta-arguments actually downplay the CMAD (i.e. they dismiss the entire debate about CMA as a waste of time, or worse). For example, it is said to be a diversion from important “efforts to target the soul of individual managers” [32]. By implication, any debate about AMA would also divert efforts “to reach the soul” of AGI/B’s (see Animism, below). A second distraction-type argument flows from the political perspective and from the “intense political struggles and scandals” that have historically surrounded corporate activities [19]. By implication, any abstract AMA debate should focus instead upon the power of AGI/B’s to control society, or vice versa. Yet another argument involving distraction is that the CMAD is premature in terms of progress in philosophy itself [25, 34]. It is claimed that the very idea of CMA (and AMA) depends on other fundamental unresolved ontological and methodological issues.

3.2 Metaphor

The concept of CMA is supported by a rich supply of metaphors between collectivities and individuals (e.g. organization-as-organism, etc.). All of these belong in the pro-CMA camp at the object-level (see sections 4 & 5) but they also conceal a problem at the meta-level with the entire CMAD: “CMA” might be nothing more than a loose way of thinking (moral projection, [17] or speaking (i.e. the translatability thesis [25]. When we talk about “corporate moral responsibilities or duties” we are being lazy or making poetical references to the ethics of individuals, just as phrase “the sun rises” refers to the Earth rotating. By implication, debates about AMA are vague references to the technical specifics of AGI/B. It has accordingly been suggested that one is “more likely to achieve good by refusing to reify abstractions” such as “corporation” or even “ethics” [21]. By implication AGI/B’s should similarly focus on technical specifics rather than abstractions like “agency” or “intelligence” per se.

3.3 Mutuality

Further meta-arguments involve subjectivity, interactivity and mutuality (i.e. philosophical pragmatism). Forty years ago Ackoff & Emery [1] argued that rationality and ethics are “in the eye of beholder”. That is, observers simply ascribe morality to any agent as they see fit, and that’s all there is to it. More recently, it has also been argued [45] in the context of the CMAD that the “only useful way to think of ‘intention’ is from a third-party perspective (cf. [12]). Classical American Pragmatism (CAP) then extends all these positions by arguing that the very meanings of words like “moral” and “agent” are never given a-priori to an inquirer. Instead, they gain their full meaning within the context of each other, as inquiry proceeds (e.g. [4]). This notion seems to be gaining ground in line with a wider rise to prominence of CAP (e.g. [42, 44]). For example CAP yields the intuitively appealing argument that the AGI/B project creates its own ethics (section 6).

4. CMA arguments

Object-level arguments for and against the very idea of CMA can also be divided into four overlapping classes (as depicted in Figure 2, above) namely: legalistic, psychological, re-casting and super-moral, as follows.

4.1 Legalistic

These are pro and anti-CMA arguments derived mainly within the law and philosophy tradition. According to the machine–metaphor [5, 40] organisations are like traditional machines controlled by individuals who must bear the moral responsibility. By implication, if an AGI is viewed as an amoral machine its builders and managers become responsible for harms caused. Against this, French [12] proposed the systems principle: that the
existence of a “corporate internal decision structure” licenses a description of corporate acts as “intentional” and hence as carrying moral responsibility. A variant known as the identification-principle holds that a layer of senior managers is the “internal structure” and the locus of responsibility. By implication, an AGI itself can be responsible because it obviously has an “internal decision structure”.

Various other considered positions suggest that moral (and legal) responsibility is generally shared or “distributed” amongst the corporate entity itself (i.e. the agency principle) and its component individuals. For example, the principle of composite-men’s-reason states that (senior) manager(s) must accept some responsibility for harm where “information acquired or developed by several different employees was never assembled into overall understanding by any one individual” [30]. Furthermore, “most cases of harm associated with corporate acts do involve personal moral fault on part of one or more individuals”. Yet it is also recognized that “often unjust and impractical to blame particular individual(s) inside a corporation”. By implication, an AGI itself ought to be punished when it has caused harm (along with any individual designers builders and managers who can be clearly shown to have “caused” the harm). This implies that an AGI has to be punishable, by, for example, slowing it down or computing a negative quasi-emotion (cf. [41]).

It has also been suggested (with reference to the “political perspective” mentioned earlier) that the power of governments (or markets) to disband entire corporations seems “totalitarian” [21]. By implication, AGI’s merit limited political rights not to be switched off or destroyed. On the other hand, corporate rights, such as the right not to be disbanded, are “conventional” rather than inalienable [29]. That is, they “depend on someone’s choice and acceptance of ‘constitutive’ rules” and are conferred by social convention. By implication a particular society might grant AGI’s “a right not to be destroyed” subject to politically determined conditions (i.e. laws established in that society). If a self-programming AGI subsequently violates those laws (i.e. post-takeoff), the “society” is then entitled to pull the plug or even to destroy it.

6 Rankin [32] has argued (on the anti-CMA side) as follows: when market power is deployed to disband a corporation (i.e. an acquisition) harm is usually caused to some individuals, whilst the target corporate-entity is disbanded; but this is not the moral equivalent of murder.

4.2 Systemic and Psychological

Several other pro and anti CMA arguments involve systems and theories of psychology, such as the claim that corporations ought to be capable of “adult” levels of moral reasoning (e.g. [39]. By implication, AGI’s also ought to be able to (i) respect other’s values, (ii) understand that it can be ethical to disobey unjust laws and (iii) invoke universal principles in moral reasoning (e.g. [41]). A second line of argument involves the concept of purposeful-systems [1]. An argument on the anti-CMA side deems that corporations are unlike humans (and by implications, not moral beings) because they contain within themselves goal-selecting sub-systems (divisions, managers, etc). On the other hand, maybe humans do too: as in the psychological concept of multiple selves [8]. The latter idea extends the individual-organization metaphor and is accordingly pro-CMA. The implications for AGI’s are simply that (i) an AGI ought to be able to manage its internal value conflicts and sub-systems, and (ii) ethics or ethical routines can be diffused throughout any system (e.g. by combining top-down and bottom-up approaches).

Other “psychological” arguments for and against CMA have involved the themes of emotion [9] character [2] and culture (i.e. the spanning-themes, as depicted in Figure 4 below). For example, corporations per se lack the “emotional make-up” that allows an individual human to show virtues and vices [9]. By implication, AGI’s need quasi-emotions in order to be able to demonstrate moral capability and have moral standing. X2 Yet other arguments involve social and cross cultural psychology. For example, the bystander effect [6] and the empirical phenomenon of “diffusion of responsibility” [22] demonstrate that individuals often tend to suppress their responses to morally relevant features of a situation when they are accompanied by others. This particular behavioural limitation adds weight to the argument that ethical routines ought to be deliberately built into organizations (and by analogy into AGI’s) in order to compensate for a likely increase in moral lapses by managers (or AGI/B’s) who are indeed typically accompanied by colleagues.

With regard to cultural tendencies, it is also noteworthy that some collectivist cultures link property rights (like those of corporate shareholders) to strong duties of care [28]. By implication, any AGI/B’s duty to care ought to be modified to reflect the relevant culture. In the last decade or so, the vexed question of how much “modification” ought to
take place seems to be tracking towards “not very much” (i.e. the absolutist/universalist position). Another argument about CMA of this type involves the role differentiation vs. role integration component of the wider business ethics debate [16]. It is generally argued that an individual ought to (temporarily) adopt the values of the relevant corporation/culture/society to some limited extent. By implication, an AGI that operates in a special context like the military should essentially adopt military codes; but only up to a point that has been settled by prior ethical deliberation and that is subject to periodic review.

4.3 Recasting and separating

A final set of CMA arguments are somewhat rhetorical. They re-cast the CMAD as part of a wider debate about ethics and strategy, which in turn has various bi-polar components such as role-differentiation vs. integration (mentioned earlier) but also value-priorities (e.g. efficiency vs. justice) or the shareholder vs. stakeholder “models” of management, and so on (Figure 3). Pro and anti-CMA arguments often “separate” the two sides of the framework (in Figure 3) by linking together several poles on the same side (e.g. justice, stakeholder model, compensation for the limitations of markets, and the pro-CMA position).

4.4 Super-moral arguments

Various CMA arguments, including several in the foregoing categories, suggest a potential for super-moral AGIs (refer back to Figure 2). Some of these arguments involve the limitations or weaknesses of individual and corporate decision making processes, implying that AGI’s can be more careful or less negligent than humans, especially when they “run” corporations (cf. [18]). The first such “weakness” argument refers to systematic biases in managerial cognitions, that is, bounded and quasi-rationality (e.g. [33]). Another limitation involves the risky-shift (or agentic-shift) in organisations [20]. By implication AGI/Bs can be much better at rational and ethical decision making (e.g. an AGI need not invoke the availability heuristic when making probabilistic predictions, nor make biased risk assessments). Yet another argument of this type involves Akrasia or the human tendency to decide upon something, but then actually do something else. Once again, AGI’s can easily be programmed to avoid this “human” foible and to exercise moral judgement independently and consistently.

Another sub-class of super-moral argument involves “extreme ethics”: behavior that seems too much to ask of humans and far beyond corporate capabilities. For example, according to Levinas “endless responsibility for the other” is the proper mark of an ethical individual, but the “hierarchical commodified relations inside a corporation” are irreconcilable with any such attitude [3]. In contrast an AGI can indeed be programmed to serve endlessly or to demonstrate a (quasi-) love ethic (i.e. artificial Agapism). It is quite clear that AGI’s are potentially superior in these respects, so they are presumably more blameworthy (along with the AG IB’s) to the extent that they fall short of this “extreme” or ideal. Finally, the notion of artificial Animism has to be considered in this context: if an AGI is programmed to believe that various other types of entity (individuals, corporations, AGI’s, animals, plants, rocks) possess souls, that AGI would presumably behave better towards the “other” and be treated well in return (cf. “interactivity” above).

5. AMA design informs CMA

As a separate matter, there are several ways in which the AGI project has indirectly informed the CMA debate (Figure 4). These all involve the circle of metaphors encompassing individuals, corporations and AGIs. Any metaphor between AGIs and corporations is subject to the critical meta-argument (section 2 above) although in the special case of the
“Virtual firm” it is an identity (as depicted in Figure 1 above). As noted elsewhere [37, 38] this metaphor can be extended with reference to several concepts and ideas developed within the AGI project, such as:

(i) The idea of easy-to-program values (e.g. survival, preserve valuable patterns, create diversity) which is new in philosophy and seems to apply also to organizations.

(ii) The concept of ethical-incrementalism, whereby ethical decision making in organizations is re-conceptualized as a series of selections of micro-actions.

(iii) Subsumptive organizational architectures (i.e. corporate department heads interact directly, with no headquarters) which were inspired by the high performance of insect-like robots where each “leg” takes its cue mainly from the other legs,

(iv) The principle that autonomy precedes sensitivity in the development of AGI’s which suggests (via the metaphor) that corporate autonomy (e.g. from shareholders) precedes “sensitivity”, and

(v) The general trend or evolution from AI towards AGI (e.g. [10]) which suggests that there is a slow but inevitably moral progress in business (e.g. [35]).

Figure 4. AGI/B informs the CMA debate

In assessing these ideas as contributions to the CMA debate (i.e. pro-CMA arguments) one should remain mindful of the meta-argument about the general limitations of metaphors. That is, although the above ideas from the AGI project seem to strengthen the case for CMA, they cannot establish particular duties or obligations for corporations per se.

6. Pragmatism

As was also mentioned earlier (section 3.3. on “mutuality”) the philosophical school of Classic American Pragmatism (CAP) has been making quite a comeback (e.g. [4, 42, 44]). This, in turn, seems at least partly attributable to progress in computer science; because there are so many obvious ways in which projects like AGI/B embrace and express the core themes of CAP, as summarised below.

(i) Inquiry per se is a major theme in CAP. Charles Pierce wrote that inquiry “originates from an irritation resulting from doubt, or because of a puzzling situation” just as AGI/Bs might be irritated or puzzled by ethics, or an AGI might detect morally-relevant features that trigger a subroutine.

(ii) Diagrammatic representations per se are seen as an important part of scientific inquiry, just as they (or programmed objects) play a vital role in AGI design and functioning.

(iii) CAP endorses comprehensive surveys of relevant facts, just as AGI/B’s might scan for all relevant data.

(iv) Goal-seeking entities are engaged in continuous invention [26] just as AGI/B’s demonstrate continuous improvement and moral-imagination

(v) CAP emphasizes the recursive nature of inquiry, which is the basis of all computing.

(vi) CAP also informs us that when we encounter ambiguities, they do not have to be regarded as roadblocks [23]. They are obstacles that can be negotiated whilst intelligent and productive activities continue.

Figure 5. The mutuality of “ethics” and “agent”

Last but certainly not least, pragmatists argue that in general “constructs only gain their full meaning and significance” within the context of each other [4]. This implies that the very meaning of “ethics” depends upon which type of “agent” or entity is being considered (Figure 5). Traditional moral philosophy contemplates the individual as a given, within its social or economic context (yielding Ethics-1 in the figure). Business and economic-ethics (ethics-2) also tends to focus on the productive activities of firms or corporations, viewed as givens. Under pragmatism, in contrast the very meanings of “ethics-1” and “ethics-2” are tied up with the nature (ontology) of “individual” and “business”. By implication, the AGI/B project co-produces its own “artificial ethics” (Ethics-3). Put differently “Robotocists are doing philosophy, whether or not they think this is so” (Dennett [7]).
7. Conclusion

The above review overall seems to suggest three broad conclusions for AGI/Bs and for further philosophical inquiry.

(i) Moral responsibilities (and hence blameworthiness for failures to live up to them) are usually distributed across all those involved: the AGI itself, its designers, builders and their managers (but also investors and other informed contributors). The precise distribution in any given case depends upon the contextual details.

(ii) AGI/B’s ought to focus upon creating well-behaved sociable AGI’s that can be reasonably expected to improve the human condition. This is an incremental process of continuous improvement. There is no need to delve into philosophical abstractions, and finally

(iii) further inquiry into artificial ethics and agency should focus more closely upon the “political meta-argument” (mentioned in section 3).

With regard to the latter, it is rather obvious that one has to consider the nature of “AGI/B” and the meaning of “ethics” in relation to the various possible political-economic arrangements in which all such projects and inquiries are embedded. For example, the military and civilian contexts are very different, but one might also consider, for example, the possibility that AGI/B’s could persuade entire electorates to adopt the laws and politics of an ideal state. Also, within business ethics in particular, there appears to be a related need to further clarify the relationships or useful distinctions (if any) between the concepts of moral-agency, identity, personhood and citizenship, as applied to corporate entities. More generally, the traditional boundaries between ethics, law, politics and technology do seem to be fading. These are not mere “perspectives” within a controversial debate; they are interrelated or mutually constituted categories of meaning.

References


[38] -------------- (2011b) Robotics as moral philosophy. *Proceedings of HICCS 44*


