The Role of Consciousness and Artificial Phronēsis in AI Ethical Reasoning

John P. Sullins

Sonoma State University, Rohnert Park CA, USA

johnsullins.com

Abstract. Phronēsis is a philosophical term that refers to conscious ethical reasoning or practical wisdom. It is argued that in most adult humans this capacity plays a primary role in high level ethical reasoning. If we want AI systems to have the capacity to reason on ethical problems in a way that is functionally equivalent to competent humans, then we will need to create machines that display phronēsis or practical wisdom in their interactions with human agents. It is argued here that this is the highest goal of AI ethics, but that this will not be a trivial problem since it may first require that the problem of artificial consciousness is solved. Achieving artificial phronēsis is necessary, since moral and ethical competence is required in order to develop ethical trust in the users of AI or robotics systems.

Keywords: Artificial Phronēsis, AI Consciousness, AI Ethics, Trust in AI, Trust in Robotics.

1 Phronēsis, Artificial and Otherwise

On one hand, consciousness is the easiest thing to experience, since consciousness is necessary for rich fully experienced perceptions, of the kind that we humans have regularly. Thus, the vast majority of those who read this paper are likely to be conscious (at least some of the time) and are fully aware of what it is like to be conscious. On the other hand, consciousness is a phenomenon that is notoriously difficult to fully explain in a way that would allow a computer engineer to build a conscious machine, even after there has been decades of heroic attempts to do so in our recent history. Since the status of consciousness studies is still in an exploratory state, I have attempted to avoid referencing it in my earlier works on AI and Robotic ethics. However, in this paper I will do so, albeit in a cautious manner.

Much work can be done in AI and Robotic ethics that does not require that the systems in question have any consciousness at all. There are three levels of artificial agents of which I want to discuss here, which are classified by their ethical abilities, or lack thereof: Ethical Impact Agents (EIA), Artificial Ethical Agents (AEA), and Artificial Moral Agents (AMA).

Ethical impact agents (EIA) need no explicit claims to consciousness to do what they do. These systems are notable only in that in their operations exhibit some autonomy and have the ability to impact human agents in ways that have ethical concern. For instance an autonomous car that, during the course of its autonomous driving operation, impacts and kills a pedestrian in an accident, (such as the accident that happened on March 18, 2018 in Tempe Arizona involving a pedestrian and a self-driving Uber car) does so completely unconsciously. The System's actions are produced by its sensors and actuators working in accordance to a program, but it has no conscious experience of the situation at hand, suffers no moral remorse after the accident, and has no emotional reaction while the event is unfolding. While we might ask plenty of questions about the safety of the autonomous car, no one blames the car itself in a conscious moral sense. That moral condemnation and all questions of legal responsibility are reserved for the emergency backup driver present in the car, and the company that built the car and is testing it on public roads.[1] Thus, there is no particular need to refer to consciousness when dealing with questions on how to ethically design and deploy EIAs.

Artificial Ethical Agents (AEA) are the next step up and differ from EIAs only in that they have explicit ethical considerations programed into their operation. Building on our earlier example, an autonomous car that was programmed to take into account some ethical calculus of value when deciding whether to risk occupants of another vehicle or its own occupants to increased risk during an otherwise unavoidable crash would be an AEA. At first look one might think that an AEA muddies the waters a bit and that the machine itself might deserve moral blame or legal responsibility, but that is just a trick of the light. The moral blame and responsibility for any adverse consequences is still fully borne by the human agents that built, deployed, licensed, and operated the vehicle. The key is that the AEA has never itself chose its own ethical standards, instead they were chosen and programed in by human agents, who therefore assume any blame or responsibility for any ethical decisions made by the system that they designed and/or deployed. The machine is not conscious of any of the events that occur based on its operations, even the ones that look to an outside party as if they were a conscious ethical choice.

It is only when we get to the level of the Artificial Moral Agent (AMA) that consciousness may play an important role. An AMA would have the ability to choose ethical behaviors that are appropriate to the situation at hand in a way that exhibits a form of practical reasoning similar to what can be seen in competent ethical reasoning found in most human agents. This means that the system either is a conscious moral agent or is functionally equivalent to one. Another way to say this is to claim that the system displays Artificial Phronēsis (AP). Of course this concept needs a lot more explanation and that is what the rest of this paper will discuss. However, at this point in the discussion we can make the conjecture that, while consciousness is not required for EIAs nor for many types of artificial ethical reasoning agents AEAs, it may play an important role in the development of much more sophisticated AMAs that would be

[1] This list is not meant to be exhaustive, political decision makers who allowed the system to use public roads should also be under scrutiny as well as many other humans who took part in the chain of events that led up to this accident.

much more useful in interacting with humans in complex social situations that tend to be bounded by shifting ethical norms.

Phronēsis is a term that many are not familiar with outside of philosophy and the word can seem a little off-putting. However, if one's goal is to create AI and robotic agents that have the capacity to reason intelligently about ethical situations, then understanding this technical term will reward those who try, given that it is so relevant to the understanding of ethics.

Phronēsis has an ancient pedigree and has come down to us largely through the tradition of virtue ethics as the skill of being able to "live well" [1]. The designers of intelligent systems do not necessarily need to become experts on the field of virtue ethics but it has been shown that some familiarity with the core concepts can enhance the design process both from the level of the designer herself and the ethical reasoning system being designed [2].

Briefly stated, phronēsis refers to the practical wisdom that a conscious moral agent uses when she is confronted with a difficult moral or ethical problem and attempts to overcome these difficulties in an intelligent manner. Given that ethical problems are always novel, no set of preconfigured answers will suffice to solve the problem, which means that learning and creativity are the hallmarks of a phronētic agent. "There is no general rule/procedure/algorithm for discerning which values, principles, norms, approaches apply; rather, these must be discerned and judged to be relevant in the first place, before we can proceed to any inferences/conclusions about what to do" [3].

AEAs might be successfully designed taking one or more ethical schools of thought into account. One might design an autonomous system that makes ethical decisions based largely on applied utilitarian or Kantian based calculations or rules, models of human moral psychology, human religious traditions, or even on the three laws of robotics developed by Isaak Asimov.[2] While one could make AEAs using any of these methods that might be useful in certain circumstances, they will all fall far short of an AMA with artificial phronēsis. "For the virtue of practical wisdom or *phronēsis* encompasses considerations of universal rationality *as well as* considerations of an irreducibly contextual, embodied, relational, and emotional nature—considerations that Kant and others have erroneously regarded as irrelevant to morality" [2].

It is understandable that systems designers will either want to ignore ethical reasoning entirely attempting to avoid the construction of even EIAs. The slightly more adventurous will attempt to apply the more computationally tractable rule based ethical systems that could result in useful AEAs. Why get involved with AMAs that require something like artificial phronēsis to work correctly? To succeed at that may require solving problems in artificial consciousness, artificial emotion, machine embodiment, etc., all of which may be computationally intractable. Let's look at what might be the reward for pursing the more difficult problem of building AMAs with artificial phronēsis.

[2] Introductory descriptions of all of these options and more can be found in [4].

1.1 The Role of Artificial Phronēsis in Automated Ethical Reasoning

Artificial Phronēsis (AP) is that claim that phronēsis, or practical wisdom, plays a primary role in high level moral reasoning and further asks the question of whether or not a functional equivalent to phronēsis is something that can be programed into machines.

If we want AI systems to have the capacity to reason on ethical problems in a way that is functionally equivalent to competent humans, then we will need to create machines that display phronēsis or practical wisdom in their interactions with human agents. This means that AP is one of the highest goals that AI ethics might achieve. Furthermore, this will not be a trivial problem since not all human agents are skilled at reasoning phronētically, so we are asking a lot from our machines if we try to program this skill into them. On top of this, the most difficult problem is that achieving AP may first require that the problem of artificial consciousness is solved, given that phronēsis seems to require conscious deliberation and action to be done correctly. Even so, the achievement of AP is necessary, since moral and ethical competence is required in order to develop ethical trust between the users of AI or robotics systems and the systems themselves. Achieving this will make for a future where humans can be comfortable inhabiting and not feel oppressed by the decisions made by autonomous systems that may impact their lives.

1.2 Artificial Phronēsis a Manifesto

AP is a new concept but it is gaining some attention. What follows are some statements to help define this new, interdisciplinary area of research.

AP Claims that phronēsis, or practical wisdom, plays a primary role in high level moral reasoning and further asks the question of whether or not a functional equivalent to phronēsis is something that can be programed into machines.

AP is a necessary capacity for creating AMAs, however the theory is agnostic on the eventuality of machines ever achieving this ability but it does claim that achieving AP is necessary for machines to be human equivalent moral agents.

AP is influenced by works in the classical ethics tradition but it is not limited to only these sources. AP is not an attempt to fully describe phronēsis as described in classical ethics. AP is not attempting to derive a full account of phronēsis in humans either at the theoretical or neurological level. However any advances in this area would be welcome help.

AP is not a claim that machines can become perfect moral agents. Moral perfection is not possible for any moral agent in the first place. Instead AP is an attempt to describe an intentionally designed computational system that interacts ethically with other human and artificial agents even in novel situations that require creative solutions.

AP is to be achieved across multiple modalities and most likely in an evolutionary machine learning fashion. AP acknowledges that machines may only be able to simulate ethical judgement for quite some time and that the danger of creating a seemingly ethical simulacrum is ever present.

This means that AP sets a very high bar to judge machine ethical reasoning and behavior against. It is an ultimate goal, but real systems will fall far short of this objective for the foreseeable future.

1.3 Dewey on the role of Phronēsis in Conscious Thought

If phronēsis was simply a concept only from ancient philosophy, it would be of limited value to the project of AI. But it has evolved over time and one very interesting development of the concept came from the philosopher John Dewey [4] [5] [6]. Unlike the ancient philosophers who seem to use phronēsis to denote a capacity that only the most highly intelligent philosophers possess, Dewey greatly expands the concept to one that plays a central role in all manner of reasoning due to the fact that once you try to apply any science or skill you necessarily enter into the social sphere and successfully operating there requires phronēsis.

If he is correct, then phronēsis is part of what makes many of us competent, conscious, and conscientious beings. It follows then that AP is either essential for the creation of conscious machines, or vice versa.

2 Ethical Trust of AI and Robotic Agents

Trust and AI and Robotic agents is its own complex topic, but here let's limit our discussion to phronētic trust. Our AP agents will need to convince us that they have our moral character in mind when they are dealing with us and will make decisions accordingly. We learn from each other how best to develop our own moral character, so these machines will need to participate in that important social process as well. In some sense they have to be able to serve as a phronēmon, or moral teacher. In these kinds of relationships with our machines, we have to be warranted in reasoning that the machine we are trusting has a good 'character' that is deserving of our trust. Without a sufficiently developed AP, then this will be impossible and there will be no good reason to try to build artificial ethical agents and we will need to limit the applications in which we employ AI to only those with no ethical impact.

2.1 How Can We Ethically Trust AI and Robotic Systems?

Mark Coeckelbergh [8] argues that there are at least two distinct ways we can look at trust in AI and Robot systems. The first is in a "contractarian-individualistic" way where we might enter into trusting relationships with these systems in a similar way in which we enter into trusting relationships with corporations or other legal entities when we rely on their products or services. These relationships may be more or less prudential but they are a long established practice in our societies. Taddeo and Floridi have developed some of the nuances that will need to be considered as these systems become more competent and the trust we place in them becomes more meaningful [9],[10]. Coeckelbergh's second category is the phenomenological-social approach. This is the kind of deep trusting relationships that humans commonly enter into with each other

that develop social bonds and create deep and meaningful relationships. Coeckelbergh is skeptical that this can be achieved in AI and robotic systems in any way that is not 'virtual-trust' or "quasi-trust" where these systems might be good at playing us in social games and garner our trust, but we do so at our own risk [8]. While we may be fooled into thinking we are in a relationship of ethical trust with the AI or Robotic system, in fact we are not, and in some situations this could be dangerous to the undiscerning human agents. Grodzinsky, Miller, and Wolf, provide a system that might mitigate this problem through the development of a new concept called "TRUST" thought of in an object oriented way in which "TRUST", "…will include traditional, face-to-face trust between humans, and "TRUST" will also include electronically mediated relationships and relationships that include artificial agents" [11]. Here we clearly define a new system of trust in machines that fits under a new heading category that also contains our already well developed notions of human to human trust and human to corporation trust, etc.

Only if we wanted machines to join us in the phenomenological-social trust would we really need AP. At that level we would join them in a new kind of society. That is a very interesting eventuality to contemplate but it is also one that would come with rights and responsibilities that would be bidirectional, so we should proceed with cation, and if it turns out that machine consciousness is technically infeasible, then it is something that we could not approach at all.

References

1. Aristotle: Nichomachean Ethics. Irwin, T (editor). Hackett, Indianapolis (1985).
2. Vallor, S.: Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting. Oxford University Press, (2016).
3. Ess, C.: Digital Media Ethics. P. 25. Polity Press, (2009).
4. Anderson, M. and Anderson, S.: Machine Ethics, Cambridge University Press, (2011).
5. Dewey, J.: The Influence of Darwinism on Philosophy. In: Hickman, L., Alexander, T. (eds.). The Essential Dewey, Volume 1. pp. 39-45. Indiana University Press (1998).
6. Dewey, J.: Evolution and Ethics. In: Hickman, L., Alexander, T. (eds.). The Essential Dewey, Volume 2. pp. 225–236. Indiana University Press (1998).
7. Rogers, M. L.: Action and Inquiry in Dewey's Philosophy. Transactions of the Charles S. Pierce Society 43(1), 90-115 (2007).
8. Coeckelbergh, M.: "Can we trust robots?" Ethics and Information Technology, 14: 53–60 (2012).
9. Taddeo, M.: Modelling Trust in Artificial Agents, a First Step Toward the Analysis of E-Trust. Minds & Machines. 20, no. 2: 243-257 (2010).
10. Taddeo, M., and Floridi, L.: "The case for E-trust," Ethics and Information Technology, 13: 1–3 (2011).
11. Grodzinsky, F., Miller, K., and Wolf, M.: Developing artificial agents worthy of trust: "Would you buy a used car from this artificial agent?" Ethics and Information Technology, Vol. 13, No. 1, 17-27 (March 2011),