Interests and Agency in AI: The case of image recognition with Inception 3 model

Completed Research Papers

Anna Sidorova
University of North Texas
Anna.sidorova@unt.edu

Abstract

The growth in AI-capabilities and proliferation of AI-enabled artifacts raises questions about unintended consequences of such technologies including the agency problems between intelligent agents and their human principals. This essay demonstrates how the agency theory and the actor-network theory (ANT) offer different, yet complementary views of the issue. Whereas the agency theory is best applied to the mitigation of the agency problem, ANT can be inform our understanding of the heterogeneous goals and interests of IT artifacts. Using an ethnographic mini-case study involving the application of machine learning algorithms to image classification, the essay traces interests inscribed in AI artifacts. The example highlights how interests of sources of training data are inscribed in AI models, and how such interests become apparent when the model is adopted by a user. Implication for future research and practice are discussed.

Keywords

Artificial intelligence, machine learning, actor-network theory, agency theory.

Introduction

Precipitous growth in artificial intelligence (AI) capabilities fueled by advances in machine learning and substantial financial backing from large IT corporations has attracted ongoing interest in the topic from the press and business IT leaders (McKinsey Global Institute 2017). The emergence of intelligent things which operate semi-autonomously or autonomously to accomplish tasks while adapting the environment is viewed by industry experts as one of the top strategic IT trends (Panetta 2017) . The growing interest in artificial intelligence (AI) on the part of business is paralleled by the increased understanding that adoption of intelligent artifacts will be accompanied by a variety of unintended and often negative consequences. While over 85% of executives count on AI to help their companies to gain or sustain a competitive advantage (Ransbotham et al. 2017), a growing number of AI researchers and practitioners draw attention to the dark side of AI. In early 2018, a group of twenty-five researchers representing the Future of Humanity Institute at University of Oxford, OpenAI, Stanford and Yale universities, as well as several other educational and research institutions, released a one-hundred page report detailing their predictions regarding the malicious use of artificial intelligence (Brundage et al. 2018). The report suggests that the fast proliferation of AI systems will be associated with the expansion of existing security threats, as more actors will be capable to carry out malicious attacks. The growth in AI capabilities will also give rise to new types of security attacks that exploit security vulnerabilities of new AI-enabled systems, such as self-driving cars or robots.

AI and IoT capabilities have been linked to the unquestionably positive outcomes, such the ability to detect malignant tumors more accurately and the improvement in treatment protocols (Drozdov et al. 2009; Ramesh et al. 2004; Verizon 2017). It is also associated with some indisputably negative consequences, such the rise in computer security threats (Brundage et al. 2018). However, the majority of AI and IoT applications are likely to be accompanied by a variety of consequences, which cannot be viewed as strictly positive or negative. Instead, the rise of intelligent artifacts is expected to precipitate

alterations in socio-technical systems through the renegotiation of contracts, redistribution of rents and shifts in the balance of power (O'Neil 2016). Predicting and managing the consequences of AI requires an understanding of how the interests of AI are aligned with the interests of other social actors. In other word, in order to understand the effects of AI it is important to understand the agency behind it. In this essay, I seek to demonstrate how the AI agency problem can be viewed differently from two theoretical perspectives, the agency theory and the actor-network theory (ANT). Specifically, the goal of the paper is to address the following research questions:

- How are interests of an AI system defined differently by the agency theory and the ANT?
- How do conflicts of interests involving AI systems emerge and are resolved when viewed from the agency theory vis-a-vis the ANT perspective?

The study contributes to AI and management research by highlighting the soci0-technical nature of AI systems and by drawing attention to the relational aspects of AI. The rest of the study proceeds with a brief discussion of intelligent agents, and a review of relevant literature on agency theory and ANT, followed by the description of the ML application mini-case and its analysis from the agency theory and ANT perspectives. The implications of this comparative analysis for AI research and practice are discussed in the conclusion.

Literature Review

Artificial intelligence

The term artificial intelligence was coined in "The Proposal for the Dartmouth Summer Research Project on Artificial Intelligence" where the scope of AI research was defined as the study of "how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves" (McCarthy et al. 1955). The conceptualization of what defines an intelligent machine can be traced to the test for machine intelligence defined by Alan Turing in terms of the machine's ability to behave like a human and to pass for a human (Turing 1950). AI research and practice to this date can be viewed as a pursuit of enabling machines to demonstrate human and above human abilities. Since its inception in the middle of the 20 century, AI research has been characterized by periods of fast progress and the so-called "winters", periods of relative inactivity following the inability of the AI community to meet the inflated expectations of the stakeholders (Russell and Norvig 2010). In early 2000s, a confluence of several technological trends led to an exponential growth in AI capabilities culminating in the achievement of several important milestones such the development of a self-driving car, as well as AI capable of winning human contestants in games such as Jeopardy, Go and Poker (Brynjolfsson and McAfee 2014:9). While most currently available systems are narrow AI designed to excel in performing specific tasks, significant progress in areas such as representation learning, transfer learning and reinforcement learning is contributing the development of Artificial General Intelligence.

Central to the field of AI is the concept of an agent, defined as "anything that can be viewed as perceiving its environment through sensors and acting upon that environment through actuators" (Russell and Norvig 2010:34). An agent is defined as rational if it selects an action most likely to maximize its performance measure given the information provided to the agent through its percepts and the prior knowledge provided to the agent by its designer. A performance measure represents the key mechanism through which the goals of the agent are defined and is typically expressed in relation to the desirability of environmental states that result from the agent's actions. An agent is considered autonomous to the extent to which it can compensate for partial or incorrect prior knowledge by learning from its actions and the percepts received from the environment. Therefore, ability to learn from its actions and from the data provided by its environment is considered a critical AI capability. Learning can be applied to different components of an agent or its environment (Russell and Norvig 2010:694). Consequently, machine learning is considered a key component of AI research and practice (McKinsey Global Institute 2017).

Agency theory

Agency theory was developed as a means for examining situations in which cooperating parties have different goals or attitudes towards work or risk (Ross 1976). It has since been applied to governance

problems, employee compensation, inter-firm contracts, etc. (Donaldson and Davis 1991; Eisenhardt 1988; Roth and O'Donnell 1996; Tosi, Gomez-Mejia, and Gomez-Mejia 1989). At the core of the theory lies the relationship between a principal (a party who delegates the work) and an agent (a party who performs the work). Most applications of agency theory are focused on resolving the agency problem, which arises "when (a) the desires or goals of the principal and agent conflict and (b) it is difficult or expensive for the principal to verify what the agent is actually doing" (Eisenhardt 1989a). Specific applications of agency theory posit that certain governance mechanisms and contractual arrangements help increase goal alignment between a principal and an agent. In addition, information systems and task characteristics are proposed to moderate the effect of different governance arrangements on the agency problem (Eisenhardt 1989a).

Actor-network theory

Actor-Network theory is a socio-materiality perspective proposed by Michael Callon and Bruno Latour in 1980s (Callon 1986; Callon and Latour 1981) and later extended by the original authors and their followers. ANT was further formalized and elaborated upon in the book *Reassembling the Social: An Introduction to Actor-network-theory* (Latour 2006). The goal of the theory is to expose the entanglement between the social and the technical by following the creation and evolution of socio-technical networks of interests surrounding the development and use of material objects (Latour 1992). In its original conceptualization, the theory focused on "actors" defined as "any element which bends space around itself, makes other elements dependent upon itself and translates their will into the language of its own" (Callon and Latour 1981, p. 286). As actors develop co-dependencies with other elements, networks of aligned interests, or actor-networks (AN), emerge.

ANT is based on the principle of relationality and views every actor, human, technical or collective, through the lens of its relationships with other actors. An actor is inseparable from its network, hence the term actor-network. External observers often experience actor-networks as coherent assemblages (such as an IT application, an organization or a human actor) and the coherence and temporal stability of these AN are taken for granted. This phenomenon is referred to as *punctualisation* (Gehl 2016). When studied more carefully, punctualized actors turn out to be a finely aligned network of individual actors, which, too, can be decomposed into networks. Modern organizations as well as IT artifacts are examples of such actor-networks (Sarker, Sarker, and Sidorova 2006).

ANT has been applied to a variety contexts including business process change (Sarker et al., 2006), IS standardization (Hanseth and Monteiro 1997; Monteiro and Hanseth 1996), and business-IT alignment (Sidorova and Kappelman 2011). Although ANT does not lend itself to an easy translation into testable hypotheses, it provides a useful lens for analyzing socio-technical processes. It suggests that sociotechnical phenomena are best understood by following actors as they seek to satisfy their interests through the process actor-networks creation. This process is referred to as translation and is detailed in (Callon 1984). The translation process is followed from the point of view of a focal actor who seeks to align the interests of other actors and actor-networks with interests of his own. The translation processes involves multiple steps such as *problematisation*, *interessement*, and *enrollment* (Callon 1984). To ensure temporal stability of the achieved alignments of interests, such interests are often inscribed into technical artifacts (e.g., a computer software or physical buildings) or other hard to change elements, such as legal contracts, or "mundane artifacts" such as a car seat belt (Latour 1992). The inscription process often requires enrollment of additional actors (such as programmers or lawyers), and therefore involves a further broadening of the network of interests. ANT does not make an *a priori* distinction between human and non-human, or between individual or collective actors, thereby making it instrumental for the study of phenomena such as development and use of information systems (Walsham 1997; Walsham and Sahay 1999). The focus of ANT on interest alignment also makes it uniquely positioned for studying the agency problem in socio-technical systems such as those involving AI.

Methodology

In order to demonstrate the socio-technical aspects of AI in use, I conducted an ethnographic mini-case study in which I documented my own interactions with an AI artifact. Case studies are commonly used as means for building and testing of theories about complex social and socio-technical phenomena, and as such, involve rigorous data collection and data analysis protocols (Eisenhardt 1989b). Here, the use of the

case study is limited to the illustration of certain theoretical concepts and is not intended as the basis for generalization. Therefore, the context of the case study was selected with simplicity and convenience in mind. Because the case documents my own interaction with AI, I opted for an ethnography as the style of inquiry (Schultze 2000). Whereas my investigation fell short of the requirement of ethnographic methodology that the investigator immerse herself into another culture (my interaction with AI took place in my own office), the investigation was ethnographic in spirit as it involved compete immersion into an activity that was new to me, and made me challenge my assumptions and engage in sense making.

The dynamics of AI research and continuous shifts in what is considered cutting edge AI makes it difficult, if not impossible, to find an artifact that encompasses all aspects of AI. Therefore, for the purpose of the investigation I chose a relatively mundane, yet prominent application of AI, namely image recognition with convolutional neural networks (CNN) (McKinsey Global Institute 2017). The goal of the interaction was to develop and test a homework exercise for a graduate level business course in applications of AI. This simple exercise involved using Tensorflow and a pre-trained CNN model, Inception 3, to recognize the contents of an image.

Image recognition with convolutional neural networks

Computer vision is an active area of AI research and a vibrant space for commercial applications of AI (McKinsey Global Institute 2017; Russell and Norvig 2010). Recent advances in the area of image and video classification were achieved due to the progress in the field of deep learning and the increased ability to train deep neural networks (DNN) (Farabet et al. 2013; LeCun, Bengio, and Hinton 2015). Although neural networks as tools for machine learning were first proposed at the very dawn of AI research, recent advances in mathematics, algorithm design and computer hardware, made it possible to train neural networks with a large number of hidden layers, each containing a large number of neurons (Géron 2017). DNN can be characterized by their architecture, with different architectures being optimized for different problem domains. Convolutional neural networks (CNN) are specifically designed for the task of image recognition, and are based on the principles underlying the functioning of a visual cortex (Hubel and Wiesel 1962, 1959). The general principle of CNN design is that the model is trained to recognize visual patterns in different parts of an image and associate such patterns with specific objects. A typical CNN consist of a number of convolutional and pooling layers, and several CNN architectures have been shown to be highly effective for image recognition, including LeNet 5, AlexNet, GoogLeNet and ResNet (Géron 2017; LeCun et al. 2015; Szegedy et al. 2015). Inception architecture of GoogLeNet was designed to perform under memory and performance budget constraints, and yet achieve very high accuracy in image classification, 3.5% top error rate (Szegedy et al. 2015).

Tensorflow is an open source software library for performing numerical computations based on data flow graphs in which nodes represent mathematical operations and edges represent multidimensional data structures called tensors (Tensorflow.com, 2018). Developed by Google engineers and open-sourced in 2015, Tensorflow is used by major corporations including Google, Twitter, SAP and Uber and is incorporated into other machine learning platforms such as IBM PowerAI (Guignard et al. 2018). Tensorflow is optimized for large-scale machine learning tasks, such as training deep neural networks, as it allows distributing computations over multiple CPUs and GPU (Abadi et al. 2016). Tensorflow libraries are accessible via a comprehensive set of Python APIs and are widely used by machine learning researchers and practitioners.

While Tensorflow offers several high level APIs for building and training CNN models, training a CNN model from scratch requires massive amounts of labeled training data, and sufficient processing resources. Therefore, a common practice is to use a pre-trained CNN model, or to re-use lower convolutional layers from a pre-trained CNN model, and retrain final layers to recognize new categories of images, a practice known as transfer learning (Géron 2017). Inception 3 model trained on data from the 2012 ImageNet Large Visual Recognition Challenge to classify images into 1000 classes is available for download from the GitHub (https://github.com/google/inception) or through the Tensorflow.com website (https://www.tensorflow.org/tutorials/image_recognition) .

What breed is my dog?

The image recognition task was based on an end-of-chapter exercises in a machine learning text (Geron 2017, p.377), and its goal was to perform image recognition on photographs of different animals using Inception 3 pre-trained convolutional neural network. The follow-up exercise involved transfer learning, i.e. re-training upper layers of the model to classify images into new classes based on the user's goals. Because one of the exercises called for at least 100 images per class, I chose my dog, a good-looking mutt of unknown origins as a photo model. The ultimate intent was to retrain the network to classify pictures into those that contain my dog and those that do not. Using an iPhone I took about 120 photos of my dog, which I transferred to my work PC in .jpg format and set out to perform the first part of the exercise, i.e. recognizing the animal depicted on the picture. After downloading and restoring Inception 3 model following directions in the book and the accompanying on-line materials, I have loaded one of the images of the dog, reshaped it appropriately, and executed the prediction function. Holding my breath, I looked at the results. The model predicted the image to be of a flatworm with 100% probability. After an emotional rollercoaster, it has finally occurred to me that the .jpg format may be the culprit. The image was promptly converted to .png and fed back to the model. The verdict returned by the model was: bluetick: 51.37%, German short-haired pointer: 12.12%, Labrador retriever: 2.67%. Other photos returned similar prediction with bluetick, a breed until then unknown to me, at the top (see Figure 1). Indeed, it was more information than I expected. I was hoping to get the image recognized as simply containing a dog.

Figure 1. Original image (left) and image recognition output (right).

Understanding the experience through the agency theory lens

From the agency theory point of view, situation described above is a rather uninteresting one. The user, myself in this case, is the principal whose goal is to recognize her dog as a dog. The agent is an instance of Inception 3 model running on the user's machine. The agent is an inanimate artifact over which the user has seemingly unlimited control. Such artifact is not controlled by anybody other than the user, and thus cannot have goals that are misaligned with those of the user. The confusion involving the image file format can be viewed as an example of information asymmetry. However, the information asymmetry is not exploited by the agent and is easily resolved by the principal, after which the agent performs to the satisfaction and in the best interests of the principal. Indeed, one may conclude that no agency problem exists in this case. Thus, although useful for designing mechanisms for mitigation of the agency problem, the agency theory is limited in its ability to identify divergent goals or interests in IT artifacts.

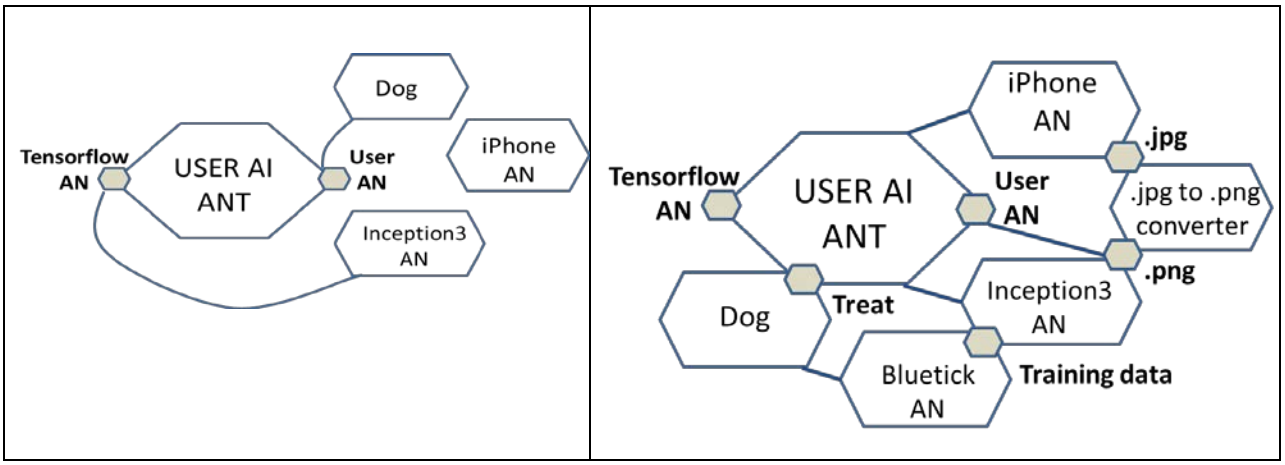
Understanding the experience through the ANT lens

ANT analysis of the user-AI interaction calls for following the principal actors as they pursue their interests. The key actor in this case is the user whose main interest is to be able to perform image

recognition. However, the user in this case cannot be separated from her larger network of interests, including her academic institution, her profession, and the artifacts she uses. Such network also includes the user's office PC with Tensorflow and the machine learning applications installed on it. Indeed, the very concept of AI and ML can be considered elements in the user's actor-network (AN), because it is in the effort of becoming an AI researcher that the user learned about Tensorflow and enrolled it into her network by installing it on her PC. Together with Tensorflow, other actors were enrolled including the Anaconda distribution platform, as well as the machine learning text by Aurelien Geron. The image classification exercise can be viewed as an effort to strengthen the user's AI AN by enrolling a new artifact, Inception 3 model.

Outside of her professional interests, the user's AN also involves her family and the dog, but the interests of the family/dog AN are relatively independent of those of her AI AN. The enrollment of the Inception 3 model required a set of 100 images that are similar to each other yet distinct from other images available to the user. Obtaining such images requires enrollment of auxiliary actors who would provide them (sources of training data). In this case, the user decides to create her own training data, which requires expanding the AI AN to include the family dog and an iPhone. The interests of a dog are easily discernable and thus its enrollment involves simple *interressment* in the form of several dog treats. An iPhone represents a complex actor-network of interests, which connects its developers, its users, the applications installed on it, and all the formats supported by it. Notably, the interests of iPhone AN remain hidden until the misalignment of such interests with those of another network is exposed. Such misalignment is uncovered when the user realizes that the default image format used by iPhone is not compatible with that of Inception 3 model. Apparently, iPhone was not designed for feeding pictures directly into Inception 3. Therefore, successful expansion of the user's AI AN to include both iPhone and Inception 3 requires enrollment of yet another actor, a program for converting .jpg images into .png format (see Figure 2).

The recognition of the probable dog breed by Inception 3 can be viewed as evidence of successful enrollment of the model into the user's AN. It can also be viewed as an enrollment of the user into the Inception 3 AN, and it reveals some of the interests inscribed in the Inception 3 model. First, by accepting images in .png format but not in .jpeg, Inception 3 promotes .png format, and with it, the interests of all actors (including artifacts) that help create .png images and to convert .jpg images into .png. By deciding to use Inception 3, the user accepts the fact that her images need to be eventually converted into .png and in the future is likely to favor data sources providing images in .png format.



Figures 2. Separate ANs before enrollment (left) and after (right).

The user's enrollment into the Inception 3 AN reveals yet another set of interests promoted by the model, perhaps unintentionally from the point of view of Inception 3 developers: the interests of dog breeding communities. All domestic dogs, *canis lupus familiaris*, have the same basic morphology which they share with gray wolves, *canis lupus* (Dewey and Bhagat 2002). However, dogs have been bread for millennia for specific behavioral characteristics and physical attributes, in order to fit the interests of particular social groups (hunters, shepherds, nobility). Viewed through the ANT lens, each dog breed represents an actor-network of inter-related interests, whereas breed names and images of dogs with representative visual

characteristics are artifacts created for inscribing such interests. By providing sufficient number of tagged images as training data for Inception 3 model, each breed actor-network ensures that its interests are represented by the model. By recognizing the image as that of a *bluetick* and not a *domestic dog*, Inception 3 model inadvertently promotes the interests of the bluetick AN. The user of the model not only became aware of the breed but also felt affinity to it, and even popularized the breed to her immediate family and friends, many of whom went on to search for pictures of the breed and read about it.

Implications for understanding agency in AI

Although the case examined here is hardly representative of the wide variety of AI use cases, the process used to highlight the socio-technical aspects of the Inception 3 model as well as the agency embedded in it, can serve as a template for understanding other AI systems and the role they can be expected to play in organizations and society. The analysis of the image recognition mini-case through the lens of the agency theory and ANT highlights several aspects of agency. Some of these are common to all IT artifacts, while others specific to AI systems.

The agency theory is primarily concerned with the goal alignment between two actors, the principal and the agent. Because inanimate objects are typically assumed not to have goals of their own, the agency theory is typically applied to relationship between human actors, such as IT departments and their outsourcing partners (Benaroch, Lichtenstein, and Fink 2016). Unlike traditional IT systems the actions of which are predefined by computer code and based on business rules established by human actors, intelligent agents are characterized by a high level of autonomy and select their actions based on a set of goals expressed in their performance function (Russell and Norvig 2010). In practice, performance measures are formulated as mathematical optimization problems and as such are not easily understood by end users of AI. And in the case of the systems based on deep learning algorithms, the action rules learned by the system are virtually impossible to interpret as they are based on hundreds or thousands of numerical weights (Géron 2017). In addition, commercial AI-enabled systems are expected to include a number of AI components for computer vision, natural language processing, policy search, etc., each operating based on its own performance function. Therefore, as the growth in AI capabilities continues, AI system will exhibit the properties of an agent whose goals are not transparent to the principal. Indeed, humans' inability to adequately set goals for AI is viewed as a serious concern in the long run (Bostrom 2012; Dubhashi and Lappin 2017). The agency theory offers insights into dealing with goal misalignment when goals of both a principal and an agent are well understood. Unfortunately, it offers little guidance in tracing the de-facto goals of an AI agent.

The ANT theory suggests methodology for following actors as they create and maintain their network of interests and helps trace the interests inscribed into an AI agent without deciphering the meaning behind mathematical notations in which the agent's program is expressed or relying on explanations provided by the AI system. Instead, this approach focusses on understanding the socio-technical alliances behind a particular AI system or its components, and relating the goals of AN to the interests of these alliances. This approach highlights the importance of sources of training data in shaping the interests of AI artifacts. Whenever an entity provides data for training AI, interests of such entity become intertwined with those of the AI.

This has important implications for developers and users of pre-trained AI models for different knowledge domains. Suppliers of training data may be able to bias the model to reflect their de-facto business practices, but also impose their preferences in data formats on future model users (Mac Namee et al. 2002). Subsequent users of pre-trained models inherit the practices inscribed in such models, and promote such practices by incorporating the models into their business processes. They are also able to influence the interests inscribed in AI by re-training it on their own business data. As such, any AI agent will carry interest traces of all actors that train it. Even when a model is trained from scratch it is influenced by the data sources that were used in the development and testing of the architecture on which it is based (Szegedy et al. 2015). As AI systems move from supervised learning based on batch data towards reinforcement learning, interests represented in training data will be increasingly difficult to identify and trace. Yet understanding the evolution of the interests of an AI artifact as it learns is very important in managing the goal alignment between intelligent artifacts and their principals.

The ANT-based analysis can be applied to a number of existing artifacts generally associated with AI. For example, success of smart speakers such as Google Home or Amazon Echo can be traced to the vast socio-technical networks with which these speakers are aligned. Google Home derives its capabilities from the network of all web sites indexed by Google and therefore, it is expected to represent the interests of those websites that more closely aligned with the Google AN. When Alexa is asked to play music, its music selection is influenced by the preferences of the vast network of Amazon shoppers. To the extent that a particular set of interests is aligned with that of Amazon AN, their interests are well represented. As AI-enabled systems become more heterogeneous, interests and goals inscribed in such systems become more diverse and less aligned with each other. Moreover, as agents given more autonomy to learn from their actions and environment, to share data and to negotiate inter-agent agreements, their interests are likely to evolve fast. Therefore, researchers and practitioners need tools and standards for tracing past associations of AI artifacts and make such information easily accessible to those seeking to employ AI as their agent.

The insights gained from the study can be useful to understanding interactions among intelligent IT artifacts. Industry experts suggest that as intelligent IT artifacts proliferate, the focus of analysis should "shift from stand-alone intelligent things to a swarm of collaborative intelligent things" (Panetta 2017). ANT is well positioned for understanding the formation and degradation of alliances of intelligent objects. It can be also applied to analyzing the impacts of technologies such as IFTTT (https://ifttt.com) and Microsoft Flow (https://flow.microsoft.com/) that are design to act as mediators between heterogeneous IT artifacts.

Limitations and directions for future research

The goal of this study is to illustrate the complex interactions and socio-technical alliances that underlie AI systems using a simple example involving the use of an image recognition model. While useful for illustrating the concepts of AI agency and interests, this example is rather simplistic. The applicability of the agency theory and ANT should be further examined through the examination of socio-technical alliances and conflicts of interests associated with complex organizational applications of AI such as robotic process automation, or fraud detection (Burgess 2018:58, 84). Because the case study methodology was used for the purpose of theory illustration, rather than theory building, it does not adhere to the strict criteria for authenticity, plausibility and criticality (Schultze 2000). Hence, future studies using more rigorous research methodology are needed to examine the issue. Furthermore, future researchers are encouraged to consider how the socio-technical alliances behind AI systems relate to specific unintended consequences such as the creation of "echo chambers" on social media personalization algorithms (Bar-Gill and Gandal 2017).

Conclusions

In conclusion, this paper demonstrates benefits and shortfalls of the Agency theory and the ANT for addressing the issue of agency in intelligent AI artifacts. Using the example of image recognition with a pre-trained CNN model, this essay demonstrates that the ANT offers a more granular view into the goals and interests carries by AI artifacts as it relates such goals and interests to current and past associations of the artifact. As such, it offers a complementary perspective to that provided by the agency theory. Further research should continue understanding AI agency issues by tracing associations between AI artifacts and socio-technical actors-networks surrounding with them. In addition, practitioners and researchers should invest in tools and standards for tracing interests and goals of AI artifacts as such artifacts are exposed to different training data and environments.

References

Abadi, Martín et al. 2016. "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems."

Bar-Gill, Sagit and Neil Gandal. 2017. "Who Gets Caught in Online Echo Chambers?" *MIT Sloan Management Review*. Retrieved May 2, 2018 (https://sloanreview.mit.edu/article/who-gets-caught-in-online-echo-chambers/).

Benaroch, Michel, Yossi Lichtenstein, and Lior Fink. 2016. "Contract Design Choices and the Balance of

Ex Ante and Ex Post Transaction Costs in Software Development Outsourcing1." *MIS Quarterly* 40(1):57–82.
Bostrom, Nick. 2012. "The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents." *Minds and Machines* 22(2):71–85.
Brundage, Miles et al. 2018. "The Malicious Use of Artificial Intelligence□: Forecasting , Prevention , and Mitigation." (February):99.
Brynjolfsson, Erik and Andrew McAfee. 2014. *The Second Machine Age□: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. W.W. Norton and Company.
Burgess, Andrew. 2018. *The Executive Guide to Artificial Intelligence*. Cham: Springer International Publishing.
Callon, Michael and Bruno Latour. 1981. "Unscrewing the Big Leviathan: How Actors Macro-Structure Reality and How Sociologists Help Them to Do so." Pp. 277–303 in *Advances in Social Theory and Methodology*, edited by In K. Knorr et A. Cicourel. Londaon: Routledge and Kegan Paul.
Callon, Michel. 1984. "Some Elements of a Sociology of Translation: Domestication of the Scallops and the Fishermen of St Brieuc Bay." *The Sociological Review* 32(1_suppl):196–233.
Callon, Michel. 1986. "The Sociology of an Actor- Network: The Case of the Electric Vehicle." *Mapping the Dynamics of Science and Technology: Sociology of Science in the Real World*.
Dewey, Tanya and Sheetal Bhagat. 2002. "ADW: Canis Lupus Familiaris: INFORMATION." Retrieved February 28, 2018 (http://animaldiversity.org/site/accounts/information/Canis_lupus_familiaris.html).
Donaldson, Lex and James H. Davis. 1991. "Stewardship Theory or Agency Theory: CEO Governance and Shareholder Returns." *Australian Journal of Management* 16(1):49–64.
Drozdov, Ignat et al. 2009. "Predicting Neuroendocrine Tumor (Carcinoid) Neoplasia Using Gene Expression Profiling and Supervised Machine Learning." *Cancer* 115(8):1638–50.
Dubhashi, Devdatt and Shalom Lappin. 2017. "AI Dangers." *Communications of the ACM* 60(2):43–45.
Eisenhardt, K. M. 1988. "Agency and Institutional Theory Explanations: The Case of Retail Sales Compensation." *Academy of Management Journal* 31(3):488–511.
Eisenhardt, Kathleen M. 1989a. "Agency Theory□: An Assessment and Review." *The Academy of Management Review* 14(1):57–74.
Eisenhardt, Kathleen M. 1989b. "Building Theories from Case Study Research." *Academy of Management Review* 14(4):532–50.
Farabet, Clement, Camille Couprie, Laurent Najman, and Yann LeCun. 2013. "Learning Hierarchical Features for Scene Labeling." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35(8):1915–29.
Gehl, Robert W. 2016. "The Politics of Punctualization and Depunctualization in the Digital Advertising Alliance." *The Communication Review* 19(1):35–54.
Géron, Aurélien. 2017. *Hands-on Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. 1st ed. Boston: O'Reilly.
Guignard, Mauricio, Marcelo Schild, Carlos S. Bederián, and Augusto J. Vega. 2018. "Performance Characterization of State-Of-The-Art Deep Learning Workloads on an IBM " Minsky " Platform." Pp. 5619–26 in *Proceedings of the 51st Hawaii International Conference on System Sciences*.
Hanseth, Ole and Eric Monteiro. 1997. "Inscribing Behaviour in Information Infrastructure Standards." *Accounting, Management and Information Technologies* 7(4):183–211.
Hubel, D. H. and T. N. Wiesel. 1962. "Receptive Fields, Binocular Interaction and Functional Architecture in the Cat's Visual Cortex." *The Journal of Physiology* 160(1):106–54.
Hubel, D. H. and T. N. Wiesel. 1959. "Receptive Fields of Single Neurones in the Cat's Striate Cortex." *The Journal of Physiology* 148(3):574–91.
Latour, B. 2006. *Reassembling the Social*.
Latour, Bruno. 1992. *Where Are the Missing Masses? The Sociology of a Few Mundane Artifacts*.
LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. "Deep Learning." *Nature* 521(7553):436–44.
McCarthy, J., M. Minsky, N. Rochester, and C. E. Shannon. 1955. *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*.
McKinsey Global Institute. 2017. *Artificial Intelligence: The New Digital Frontier?*
Monteiro, Eric and Ole Hanseth. 1996. "Social Shaping of Information Infrastructure: On Being Specific about the Technology." Pp. 325–43 in. Springer, Boston, MA.
Mac Namee, B., P. Cunningham, S. Byrne, and O. I. Corrigan. 2002. "The Problem of Bias in Training

Data in Regression Problems in Medical Decision Support." *Artificial Intelligence in Medicine* 24(1):51–70.

O'Neil, Cathy. 2016. *Weapons of Math Destruction□: How Big Data Increases Inequality and Threatens Democracy*. Penguin Books Ltd.

Panetta, Kasey. 2017. "Gartner Top 10 Strategic Technology Trends for 2018 - Smarter With Gartner." *Gartnerr*. Retrieved March 19, 2018 (https://www.gartner.com/smarterwithgartner/gartner-top-10-strategic-technology-trends-for-2018/).

Ramesh, A. N., C. Kambhampati, J. R. T. Monson, and P. J. Drew. 2004. "Artificial Intelligence in Medicine." *Annals of the Royal College of Surgeons of England* 86(5):334–38.

Ransbotham, Sam, David Kiron, Philipp Gerbert, and Martin Reeves. 2017. "Reshaping Business With Artificial Intelligence." *MIT Sloan Management Review*.

Ross, Stephen A. 1976. "The Economic Theory of Agency: The Principal's Problem." *The American Economic Review* 63:134–39.

Roth, K. and S. O'Donnell. 1996. "FOREIGN SUBSIDIARY COMPENSATION STRATEGY: AN AGENCY THEORY PERSPECTIVE." *Academy of Management Journal* 39(3):678–703.

Russell, Stuart and peter Norvig. 2010. *Artificial Intelligence: A Modern Approach, 3rd Edition | Pearson*. 3rd ed. Pearson.

Sarker, Suprateek, Saonee Sarker, and Anna Sidorova. 2006. "Failure□: An Actor-Network Perspective." 23(1):51–86.

Schultze, Ulrike. 2000. "A Confessional Account of an Ethnography about Knowledge Work." *MIS Quarterly* 24(1):3–41.

Sidorova, Anna and Leon Kappelman. 2011. "Better Business-IT Alignment through Enterprise Architecture: An Actor-Network Theory Perspective." *Journal of Enterprise Architecture* 39–47.

Szegedy, Christian, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. "Rethinking the Inception Architecture for Computer Vision."

Tosi, Henry L., Luis R. Gomez-Mejia, and Luis R. Gomez-Mejia. 1989. "The Decoupling of CEO Pay and Performance: An Agency Theory Perspective." *Administrative Science Quarterly* 34(2):169.

Turing, Alan M. 1950. "Computing Machinery and Intelligence." *Mind. A Quarterly Review of Psychology and Philosophy* LIX(236):433–60.

Verizon. 2017. *State of the Market: Internet of Things 2017 Making Way for the Enterprise.*

Walsham, G. 1997. "Actor-Network Theory and IS Research: Current Status and Future Prospects." Pp. 466–80 in *Information Systems and Qualitative Research*. Boston, MA: Springer US.

Walsham, Geoff and Sundeep Sahay. 1999. "GIS for District-Level Administration in India: Problems and Opportunities." *MIS Quarterly* 23(1):39.